FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)

GEI - MENCIÓ EN COMPUTACIÓ

# *MACHINE LEARNING* PER AL RECONEIXEMENT I CLASSIFICACIÓ D'ENTITATS MÈDIQUES EN INFORMES CLÍNICS

Projecte elaborat per Miguel Ángel Merino Santiago per la Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

---

Director:
Lluis Padro Cirera

Codirector:
Jordi Turmo Borrás

Departament de Ciències de la Computació

Abril de 2020

## Resum

El reconeixement i classificació d'entitats mèdiques és un problema que es troba a l'ordre del dia dins el camp del processament del llenguatge natural (NLP). Resoldre'l de la millor manera possible pot tenir un impacte social molt positiu. L'objectiu d'aquest projecte serà desenvolupar dos models diferents per tractar de resoldre aquesta problemàtica, el primer d'ells basat en CRFs i el segon en xarxes neuronals artificials. Els resultats ens confirmen que el segon mètode té un millor comportament, obtenint uns resultats força positius, i aportant informació per a futures noves línies d'investigació.

## Resumen

El reconocimiento y clasificación de entidades médicas es un problema que se encuentra a la orden del día dentro del campo del procesamiento del lenguaje natural (NLP). Resolverlo de la mejor manera posible puede tener un impacto social muy positivo. El objetivo de este proyecto será desarrollar dos modelos diferentes para tratar de resolver esta problemática: el primero de ellos basado en CRFs y el segundo en redes neuronales artificiales. Los resultados nos confirman que el segundo método tiene un mejor comportamiento, obteniendo unos resultados bastante positivos, i aportando información para futuras líneas de investigación.

## Abstract

The recognition and classification of medical entities is a relevant problem currently in the field of natural language processing (NLP). Solving it in the best possible way can have a significative positive impact on the society. The goal of this project is the development of two different models to solve this problem: the first one uses CRFs, and the second one is based on artificial neural networks. The results obtained prove that the second method has a better behaviour, obtaining quite positive results and bringing information to future investigation lines.

# Índex

# Índex de figures

# Índex de taules

# 1. Abast i contextualització

## 1.1. Introducció

Al llarg de la història, s'han donat diferents moments en els que el descobriment o avanç en determinades àrees ha tingut un impacte social molt important, canviant la manera de viure de les persones.

Als darrers anys, tots els avenços tecnològics han causat que ens trobem en un d'aquests moments de la història. És innegable que les noves tecnologies ja porten anys canviat la manera de concebre molts aspectes socials, i és previsible que aquest desenvolupament tecnològic segueixi avançant a un gran ritme.

Si bé és cert que la discussió de si l'impacte és positiu o negatiu es troba sempre oberta, és el nostre deure com a Enginyers Informàtics tractar d'aplicar els nostres coneixements i dedicació en aconseguir que aquest impacte sigui positiu per la gran majoria de la societat.

Aquest projecte es centrarà en el sector sanitari, un sector que tot i haver avançat molt lligat a aquest desenvolupament tecnològic, segueix sent el focus de moltes problemàtiques i conflicte social.

La voluntat serà contribuir a millorar aquest sector, tractant d'aplicar tècniques innovadores pròpies d'aquest nou mon digital, amb l'objectiu de millorar diferents processos dins el sistema sanitari, tot contribuint així a millorar la qualitat i l'agilitat d'aquest.

En concret, s'aplicaran tècniques d'intel·ligència artificial per a poder recopilar i tractar de manera automàtica la informació que es troba als informes clínics escrits pels metges, podent manipular més endavant aquesta informació obtenint així múltiples avantatges.

## 1.2. Contextualització

### 1.2.1. Projecte dins el marc de la FIB

El desenvolupament d'aquest treball d'investigació, s'emmarca dins el projecte PROSAMED (*Procesamiento semántico textual avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes Médicos*)[1] que té la voluntat d'elaborar diferents eines automàtiques d'anàlisi textual per al tractament d'informes clínics hospitalaris i història clínica electrònica, millorant processos que avui en dia tenen un gran cost personal i econòmic i donant així un gran salt tecnològic dins d'aquest sector.

Aquest projecte, finançat pel Ministerio de Economia y Competitividad i amb codi TIN2016-77820-C3-3-R, el coordina el grup d'investigació IXA de la Universidad del País Vasco (UPV) i a més, participen dos altres equips: el NPL IR GROUP de la UNED i el *Center for Language and Speech Technologies and Applications* (TALP) de la UPC.

El TALP és un grup de recerca interdepartamental especialitzat en el processament del llenguatge natural tant escrit com parlat, que actualment treballen en el projecte GRAPH-MED[2] (situat dins PROSAMED), que és on es desenvolupa aquest TFG.

### 1.2.2. Termes i conceptes

A continuació, s'exposen alguns termes i conceptes que cal entendre ja que formen una part important dins d'aquest estudi:

- **Processament del llenguatge natural (en anglès, NLP)**: és un camp d'investigació que mescla ciències de la computació, intel·ligència artificial i lingüística per estudiar i tractar les interaccions entre les computadores i el llenguatge humà.
- **Reconeixement d'entitats nombrades (en anglès, NER)**: és una tasca que trobem dins el NLP que es basa en la localització i classificació d'entitats en categories predefinides prèviament.
- **Aprenentatge supervisat**: Conjunt de tècniques del *Machine Learning* caracteritzades per utilitzar dades d'entrada correctament etiquetades i entrenar a partir d'aquestes.

### 1.2.3. Problema a resoldre

El problema que volem resoldre es classifica com a problema de NLP ja que les eines que volem desenvolupar treballen directament sobre textos escrits en llenguatge natural i això fa que aquest sigui el punt principal d'estudi.

L'objectiu, serà reconèixer entitats mèdiques i classificar-les segons unes categories preestablertes amb anterioritat. Això és, per tant, una tasca de Reconeixement d'Entitats Nombrades (NER).

Per fer-ho, treballarem sobre el corpus de dades d'informes clínics aportat per la Fundació IDIAP Jordi Gol[1].

Tot i que el problema que aborda PROSAMED considera que una mateixa entitat pugui estar associada a diverses categories, aquest abast el reduirem al nostre problema, centrant-nos en l'assignació d'una categoria per entitat.

Com s'ha especificat, la voluntat es classificar aquestes entitats en unes categories preestablertes que seran les següents:

- Diagnòstics
- Fàrmacs
- Parts del cos
- Signes / Símptomes

Un aspecte rellevant a considerar sobre el problema que volem resoldre és que el llenguatge utilitzat pels metges quan redacten els informes clínics té moltes peculiaritats que més endavant comentarem, i que fa que haguem de dissenyar un tractament especial per encarar aquesta part del problema.

La tasca, la podem veure representada en exemples a la Figura 1.1, on es fa èmfasi en mostrar algunes de les característiques del textos que ens trobarem, com faltes d'ortografia, acrònims i abreviatures:

[1] `https://www.idiapjgol.org/index.php/ca/`

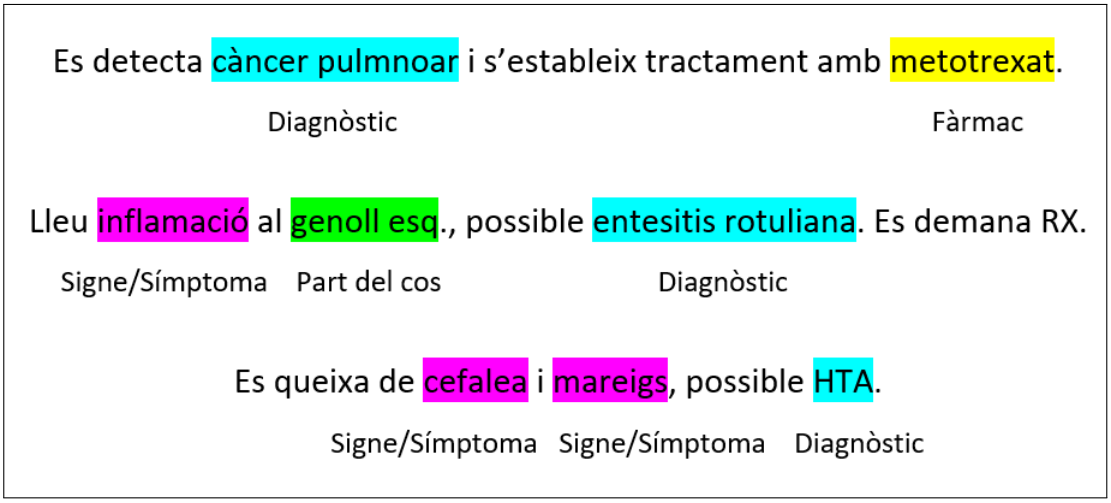


Figura 1.1: *Exemple de reconeixement i classificació d'entitats*

### 1.2.4. Estat de l'art

Les tasques de reconeixement i classificació d'entitats dins un domini mèdic s'enfoquen avui en dia amb un ampli ventall de tècniques cercant els millors resultats possibles.

Si ens centrem en alguns projectes de referència en llengua anglesa com poden ser el ib2b-2010[4] o el SemEval-2013[5], on s'arriba a resultats[2] propers o fins i tot superiors al 0.9 de rati d'encert, veiem que les tècniques es basen en el *Machine Learning*.

Els millors resultats que s'aconsegueixen, ho fan utilitzant tècniques d'aprenentatge supervisat com els CRFs (*Conditional Random Fields*) o SVMs (*Support-vector machines*).

Seguidament, si ens centrem en els projectes escrits en llengua castellana com poden ser el eHealth-KD Challenge 2019[6], pertanyent al IberLEF-2019, viem que els ratis d'encert ronden generalment el 0.60 - 0-65. No obstant, es destacable la participació del TALP, aconseguint un resultat en el millor dels escenaris d'un 0.82.

En aquest cas, les tècniques utilitzades s'enfoquen més cap a l'ús de xarxes neuronals basades en arquitectures com LSTM (*Long short-term memory*) o CNN (*Convolutional neural network*).

[2]Els resultats exposats fan referència a la puntuació F1

En qualsevol cas, tot i que observem una considerable amplada pel que fa al conjunt de tècniques que s'apliquen, el que queda clar és que totes elles pertanyen al camp del *Machine Learning*, que és on ens enfocarem nosaltres.

### 1.2.5. *Stakeholders*

La finalitat d'aquest projecte no és pas elaborar un producte comercial, sinó que com a treball d'investigació, la nostra voluntat és elaborar noves tècniques que produeixin els millors resultats possibles i que alhora contribueixin a l'avanç i innovació dins el camp en el que treballem.

Aquesta investigació, va enfocada a la futura obtenció d'un producte que pugui ser utilitzat pel sector sanitari, tan privat com públic. Aquest sector es veuria en gran mesura beneficiat perquè aquesta identificació i classificació d'entitats mèdiques es passaria a fer de forma automàtica, de manera que es disposaria de les dades amb rapidesa, sense necessitat de destinar una gran quantitat de recursos humans en fer aquest processament.

Els beneficiats, a part dels empleats del sector sanitari com hem comentat, seran clarament els propis usuaris, ja que la implementació d'aquests nous mètodes desencadenarà un alliberament de recursos que agilitzarà tot el sistema sanitari.

A més, el tractament d'aquesta gran quantitat de dades permetrà poder analitzar-les més a fons i amb major qualitat, permetent així observar noves tendències i aportant més informació al sector d'investigació clínic.

Per tant, aquest treball es una petita part o contribució a un projecte més gran que vol beneficiar en gran mesura a diferents grups vinculats amb el sector sanitari.

## 1.3. Justificació

Com s'ha mencionat anteriorment, el camp del reconeixement i classificació d'entitats dins una ontologia mèdica es tracta majoritàriament en projectes enfocats a la llengua anglesa.

Pel que fa als projectes desenvolupats en llengua castellana, és important comentar que són molts menys i que aquests encara no assoleixen uns resultats tant positius com els que es fan per l'anglès. De fet, la quantitat de projectes decreix encara més si ens enfoquem en la llengua catalana.

És per això que aquest treball vol contribuir en avançar aquestes investigacions en la llengua castellana i especialment en el català.

D'altra banda, el sector sanitari és un sector que es troba constantment en el focus d'atenció per la seva importància en el benestar de la població. Generalment, s'acostuma argumentar una manca d'agilitat i rapidesa d'aquest (en especial el sector públic).

Amb això en ment, el que busquem és contribuir a la resolució d'aquesta problemàtica que realment té un impacte social molt gran i que pot aportar molts beneficis a la població.

L'objectiu és reduir el cost personal que té aquesta cerca no automatitzada en informes clínics, tractant d'automatitzar-la i de millorar els resultats obtinguts.

S'espera que si s'aconsegueix això, aquests recursos personals es poguessin dedicar a altres tasques del sistema, tot desencadenant així una gran millora en la qualitat del sector sanitari.

## 1.4. Abast

### 1.4.1. Objectiu del projecte

En base al que s'ha explicat fins a aquest punt, acotem els objectius d'aquest projecte i els enumerem de la següent manera:

- Implementar dos models basats en tècniques diferents per a dur a terme la tasca de reconeixement i classificació d'entitats mèdiques.
- Experimentar amb aquests models amb l'objectiu d'obtenir uns resultats positius, fent un bon anàlisi i comparativa entre els dos sistemes implementats.
- Enfocar el disseny d'aquests models de manera que els resultats puguin contribuir a aportar nova informació dins els avenços en aquest camp.

És important mencionar que abans de començar a desenvolupar els models, serà necessari un anàlisi del corpus de dades del que disposem, amb l'objectiu de poder extreure'n les característiques més importants.

### 1.4.2. Possibles obstacles i riscos

Com a tot projecte, són múltiples els obstacles que poden sorgir al llarg del seu transcurs. Si bé és cert que sempre en poden aparèixer alguns que no hagin estat contemplats prèviament, tractar de localitzar tots el possibles riscos als que ens exposem pot fer que ho puguem contemplar a l'hora de considerar càrregues de treball i previsions temporals.

La majoria d'aquests obstacles, van associats a la incertesa tecnològica pròpia d'aquest treball on es tracten d'utilitzar tècniques i mètodes innovadors. Això vol dir que durant l'elaboració del projecte, aniran sorgint problemes derivats de les tècniques usades que caldrà resoldre.

D'altra banda, és important considerar que el fet d'estar implementant uns mètodes d'aprenentatge supervisat, necessitarem que les dades sobre les que treballem tinguin uns mínims de qualitat, realitzant sobre elles les correccions necessàries quan calgui.

### 1.4.3. Metodologia i rigor

La decisió sobre quines metodologies emprar durant el desenvolupament del projecte és clau, ja que una decisió dolenta pot desencadenar un progrés lent i amb entrebancs, mentre que la elecció de les metodologies de treball adients millorarà notablement el rendiment i eficàcia a l'hora de dur a terme el projecte, millorant la qualitat d'aquest.

En primer lloc, s'ha contemplat la opció d'utilitzar una metodologia de desenvolupament en cascada, on es segueixen de manera seqüencial les següents etapes: requisits, disseny, implementació, verificació i manteniment.

No obstant, pensar en utilitzar aquesta metodologia dins el nostre projecte de recerca no casava del tot, ja que el fet d'haver de finalitzar una etapa per passar a la següent, en un projecte que tracta un àmbit tant innovador, pot ser clarament un inconvenient. Per exemple, es pot donar la situació en la que posteriorment a la implementació i testeig d'un mètode, ens adonem que els resultats obtinguts no són del tot els esperats i decidim modificar el disseny del propi mètode. Això no seria vàlid en una metodologia de treball en cascada, de fet, podem veure que el que té més sentit és algun tipus de metodologia de caire més iteratiu.

És per això, que la metodologia de treball escollida finalment és el sistema iteratiu i incremental. Aquest sistema consta d'una etapa d'inicialització i posteriorment una etapa iterativa. Aquesta etapa iterativa el que ens permet és realitzar cicles on a cada un el que fem és: dissenyem un mètode, l'implementem, observem resultats i considerem si és necessari redissenyar el mètode. De ser així, tornaríem a realitzar un altre cicle.

D'aquesta manera, ens beneficiem de l'aprenentatge que obtenim a cada iteració, ja que podem utilitzar-lo a la iteració posterior amb la finalitat d'obtenir millors resultats. Aquest aspecte és molt positiu tractant-se d'un projecte d'innovació, i clarament casa molt millor amb el desenvolupament que volem seguir.

D'altra banda, amb la voluntat de mantenir una visualització més dinàmica de l'estat del projecte per a poder tenir un millor seguiment, s'ha optat per fer ús d'alguna metodologia *agile*. En aquest cas, s'ha escollit el Kanban, mantenint una taula on s'agruparan les tasques en funció de si estan pendents de fer, en progrés o finalitzades. Gràcies a això podrem visualitzar tot el flux de treball, tenint clar la feina en curs que hi ha a cada moment, i que és el que s'ha de prioritzar per no bloquejar altres tasques.

Emprant aquestes metodologies exposades anteriorment, el que s'espera és optimitzar les hores de treball seguint una elaboració estructurada i òptima del projecte i aconseguint al final un projecte sòlid i de qualitat.

Finalment, pel que fa als mètodes de validació d'objectius amb els directors del projecte, es realitzarà una sèrie de reunions físiques amb la freqüència adient per tal d'assegurar un correcte desenvolupament del treball. Comentar que a causa de la situació excepcional que es viu actualment a causa del Covid-19, les reunions que inicialment es feien presencialment, s'han passat a fer via telemàtica.

A més, per comentar coses sobre el treball d'una manera més dinàmica, s'utilitzarà el correu electrònic sempre i quan el dubte en qüestió pugui ser tractat per aquest canal.

# 2. Planificació temporal

## 2.1. Planificació i desviacions

En el moment d'afrontar projectes de gran dimensió, és molt important dedicar un temps previ a analitzar el conjunt de tasques que conformen aquest treball i tenir una aproximació més o menys real del total d'hores que s'haurà de dedicar a cada una.

A més, en aquesta secció comentarem els canvis que s'han produït en vers a la planificació que es va realitzar inicialment, tot justificant-los per a poder seguir amb la planificació definitiva.

Per començar, tot i que el projecte es va iniciar al setembre del quadrimestre de tardor 2019/2020 i inicialment s'esperava entregar al torn de Gener, finalment s'-ha prorrogat fins al torn d'Abril del quadrimestre de primavera d'aquest mateix curs.

D'altra banda, la planificació inicial dels objectius del projecte també ha patit modificacions. Inicialment, s'esperava disposar de la informació necessària al corpus de dades per a poder classificar les entitats segons el seu codi mèdic[1]. Finalment, aquest document no ha pogut arribar a temps i per tant s'ha decidit classificar les entitats com s'ha mencionat al capítol anterior, traslladant l'esforç a tractar de millorar la implementació i anàlisi dels mètodes.

Així doncs, donat que la lectura del projecte serà efectiva al torn d'Abril (que comença el dia 27 d'aquest mes), es marcarà com a data de finalització el dia 19 d'Abril, una setmana abans de que comenci el torn de lectura.

La planificació temporal considera un total de 500 hores de treball, que distribuïdes al llarg de aproximadament 4 mesos, resulta una càrrega de treball de 4 hores diàries.

Partint de les bases comentades, a continuació s'exposarà en detall la planificació temporal que es segueix fins a la data de finalització establerta.

[1]Les codificacions oficials que es volien utilitzar eren CIE-10 (Clasificación internacional de enfermedades) elaborada per la OMS i CIAP-2 (*Classificació Internacional d'Atenció Primaria*), creada per la Oxford University.

## 2.2. Descripció de les tasques

A continuació, s'exposarà cada tasca de manera detallada, definint per cadascuna no només en què consisteix, sinó també les hores que s'espera dedicar i les dependències que té amb les altres.

La seqüencia lògica de realització de totes les feines és la mateixa que l'ordre en el que s'exposen a continuació (a excepció de les parts de documentació). L'única cosa a mencionar és que part d'aquestes tasques es podran realitzar en paral·lel ja que no suposes dependència entre elles. Aquest paral·lelisme entre les tasques quedarà es podrà observar a la secció dedicada al diagrama de *Gantt*.

Exposem la distribució de les tasques amb un resum, temps que s'espera dedicar i dependències:

- **T1 - Gestió del projecte** [**20h**]: Inclou les tasques relacionades amb la planificació temporal, la gestió de riscos i obstacles, l'informe de sostenibilitat i les metodologies de gestió de projectes utilitzades.
- **T2 - Seguiment del projecte** [**20h**]: Inclou totes les reunions i temps dedicat al control i seguiment del projecte.
- **D1 - Documentació inicial** [**30h**]: Introducció, context i justificació del projecte. Explicació dels mètodes de gestió utilitzats. Tota la part inicial de la documentació.
- **D2 - Documentació de mètodes** [**30h**]: Part de la documentació relacionada amb l'explicació dels mètodes implementats. Depèn de les T4 i T5 perquè necessitem que estiguin implementades.
- **D3 - Documentació de resultats i conclusions** [**30h**]: Part final de la documentació on s'exposen resultats i conclusions. Depèn de la tasca T8.
- **T3 - Anàlisi mètodes existents pel reconeixement i classificació d'entitats** [**40h**]: Estudi a fons dels mètodes existents i quins implementar per la tasca del reconeixement i classificació d'entitats.
- **T4 - Implementació del mètode 1** [**70h**]: S'implementa el primer mètode de reconeixement i classificació. Depèn de la T3 per la necessitat de saber quin mètode implementar.
- **T5 - Implementació del mètode 2** [**70h**]: S'implementa el segon mètode de reconeixement i classificació. Depèn de la T3 per la necessitat de saber quin mètode implementar.
- **T6 - Testing i millores mètode 1** [**70h**]: Es realitzen les proves i millores necessàries pel primer mètode. Depèn de la T4 per la necessitat de tenir-lo implementat.

- **T7 - Testing i millores mètode 2 [70h]**: Es realitzen les proves i millores necessàries pel segon mètode. Depèn de la T5 per la necessitat de tenir-lo implementat.
- **T8 - Anàlisi de resultats [50h]**: Es comparen a fons els mètodes i resultats obtinguts. Per cada mètode s'estudia a fons el seu comportament i la justificació dels resultats que ha generat. Depèn de les tasques T6 i T7 per la necessitat de tenir les implementacions definitives per a cada mètode.

| Codi | Nom descriptiu | Temps | Dependències |
|---|---|---|---|
| T1 | Gestió del projecte | 20h | - |
| T2 | Seguiment del projecte | 20h | - |
| D1 | Doc. inicial | 30h | - |
| D2 | Doc. mètodes | 30h | T4, T5 |
| D3 | Doc. resultats i conclusions | 30h | T8 |
| T3 | Anàlisi mètodes existents | 40h | - |
| T4 | Implementació mètode 1 | 70h | T3 |
| T5 | Implementació mètode 2 | 70h | T3 |
| T6 | Test i millores mètode 1 | 70h | T4 |
| T7 | Test i millores mètode 2 | 70h | T5 |
| T8 | Anàlisi de resultats | 50h | T6, T7 |

Taula 2.1: *Resum de les tasques.*

## 2.3. Estimacions i *Gantt*

Una bona manera de representar de manera gràfica el temps previst per a la dedicació en diverses tasques i la possible concurrència d'aquestes, és el diagrama de *Gantt*. A la Figura 2.1, s'observa el diagrama fet per a les tasques exposades a la secció anterior.

Com podem veure, les tasques es diferencien en colors seguint el següent patró:

- **Vermell**: Tasques relacionades amb la gestió i seguiment del projecte i que conseqüentment es trobaran en actiu al llarg de tot el treball.
- **Blau**: Feines relacionades amb la secció del TFG dedicada als mètodes de reconeixement i classificació d'entitats mèdiques.
- **Verd**: Tota aquella feina referida a la elaboració de la documentació.

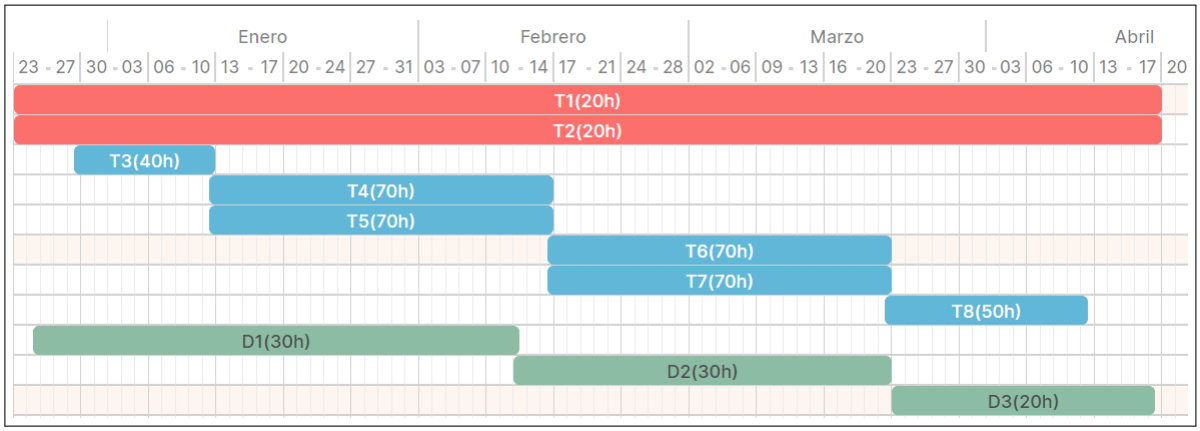


Figura 2.1: *Diagrama de Gantt*

## 2.4. Recursos humans i materials

Com és lògic, el desenvolupament del projecte requereix de recursos tan de caire humà com material.

Pel que fa als recursos de la vessant humana, la gran majoria provenen del propi estudiant que realitza el TFG. No obstant, també són requerits altres recursos humans en relació al control del projecte, com son el director i codirector del treball pels recursos destinats al control i suport de la branca més tècnica del projecte.

D'altra banda, també es requereixen un seguit de recursos informàtics. En primer lloc, llistem aquells recursos de *hardware*:

- Portàtil personal (Intel Core i5-4210U 1.70GHz, 4GB memòria RAM).
- HPC (*High Performance Computer*) del rdlab[2] del Departament de Ciències de Computació de la UPC (accés a 20 CPUs i emmagatzematge).

[2] `rdlab.cs.upc.edu/hpc/`

Per finalitzar, enumerem els recursos informàtic utilitzats en la vessant de *software*:

- Linux (Ubuntu 18.04.2 LST) com a sistema operatiu sobre el que s'ha treballat.
- Les principals llibreries utilitzades de Python han estat: Scikit-learn i Keras (executada sobre TensorFlow).
- LaTeX per a l'elaboració de tota la documentació.

## 2.5. Gestió del risc: plans alternatius i obstacles

Per finalitzar aquest capítol, en aquesta secció (que va estretament lligada a l'apartat *1.4.2 Possibles obstacles i riscos*), volem fer èmfasi en quines alternatives es consideren als possibles riscos detectats.

El principal obstacle que hem contemplat és la incertesa tecnològica que comporta treballar amb tècniques i mètodes innovadors que en alguns aspectes poden tenir mancances en documentació o suport. Per a fer front a aquesta possible complicació, el que s'ha fet és una sobreestimació general per les tasques relacionades amb la implementació d'aquests mètodes.

En concret, s'ha fet una sobreestimació de 10h tant a la tasca T4 com a la tasca T5 (tot i que ja es contempla que són tasques que tenen com a component principal el fet de fer front a desenvolupar les noves tècniques, la sobreestimació és per si suposés un esforç és rellevant de l'esperat).

D'altra banda, per afrontar els possibles problemes derivats d'assegurar una bona qualitat de les dades, es fa una sobreestimació de 5h tant a la tasca T6 com T7, que és on s'espera trobar aquestes possibles complicacions.

# 3. Gestió econòmica, lleis i regulacions

## 3.1. Pressupost

El pressupost d'aquest projecte queda definit a la Figura 3.1. No obstant, a continuació s'anirà exposant de manera detallada per definir amb exactitud la seva elaboració i per poder justificar-lo correctament.

En primer lloc, esmentar que s'ha dividit l'informe del pressupost en els següents quatre punts:

- **Costos per cada tasca o activitat** (associats a costos de personal)
- **Costos genèrics**
- **Contingència**
- **Partida d'imprevistos**

Pel que fa als costos d'activitat, s'ha associat un cost a cada tasca exposada al diagrama de *Gantt* mostrat al capítol anterior. Per a poder determinar el cost de cada activitat, s'ha d'estimar un cost per hora de treball de la persona encarregada d'aquestes tasques.

En aquest cas, i en base a un informe elaborat per Adecco l'any 2018[1], s'estimarà que el preu per hora d'un enginyer informàtic junior serà de 11 euros. A més, a l'hora de tenir en compte aquest cost, es multiplicarà per 1.35 per afegir el cost de la seguretat social, fet que resulta en un cost per hora de treball de 14.85 euros. Per tant, aquest cost i el número d'hores per tasca serà el que determinarà el cost de cadascuna.

Seguidament, es calculen els costos genèrics. Per fer aquest càlcul, aquests costos es dividiran en dos partides: la primera d'ella dedicada a costos purament genèrics que inclourà material, electricitat i desplaçament, i una segona partida dedicada als costos d'amortitzacions del HW i SW utilitzats, que seran els mencionats a l'apartat *2.4 Recursos humans i materials*.

[1] `infoempleo.com/informe-infoempleo-adecco`

| Activitat | Import (€) | Descripció |
|---|---|---|
| T1 | 297 | |
| T2 | 297 | |
| T3 | 594 | Tasques que apareixen al diagrama de *Gantt* |
| T4 | 1039,5 | |
| T5 | 1039,5 | (L'estimació del cost es fa multiplicant la |
| T6 | 1039,5 | quantitat d'hores estimades al *Gantt* per |
| T7 | 1039,5 | el preu per hora estimat de treball (14.85€) ) |
| T8 | 742,5 | |
| D1 | 445,5 | |
| D2 | 445,5 | |
| D3 | 445,5 | |
| Total CPA | 7425 | Total costos de personal per les activitats del Gantt |
| Cost genèric | 300 | Material, electricitat, desplaçaments... |
| Amortitzacions | 1500 | Ús de Software i Hardware previament adquirit |
| Total CG | 1800 | Total costos imputats genèricament |
| Total Costes (CPA + CG) | 9225 | Total costos |
| Contingència | 1383,75 | Rati establert del 15% sobre el total de costos |
| Total CD+CI+Contingència | 10608,75 | |
| Sobreestimació T4 i T5 | 89,1 | Sobreestimació de 10h per T4 i 10h per T5 |
| Sobreestimació T6 i T7 | 44,55 | Sobreestimació de 5h per T6 i 5h per T7 |
| Total imprevistos | 133,65 | (No sumem al tota, ja ho considerem previament) |
| **TOTAL:** | **10608,75** | **Pressupost total** |

Figura 3.1: *Pressupost del projecte*

Per realitzar el càlcul de les amortitzacions, hem emprat la següent formula genèrica per les amortitzacions de hardware: Cost de l'equip / (4 anys de vida útil * 220 dies feiners/any * hores de dedicació/dia) * hores dedicació TFG.

Tenint en compte que considerem un cost total de l'equip de 10.560 euros i que recordem que la dedicació total al TFG és de 500h, obtenim un cost d'amortitzacions de 1500 euros.

Sobre el total de costos elaborats fins ara, haurem d'aplicar els costos de contingència, associats a possibles sobrecostos o obstacles no considerats. En quest cas i donat que ens trobem en el sector informàtic, establirem la contingència en un rati del 15%. Si aquests costos de contingència no es produïssin, els diners d'aquesta partida es podrien utilitzar per a possibles imprevistos no contemplats, o per a futurs projectes dins aquesta àrea.

Per finalitzar, comentar que pels possibles imprevistos identificats del projecte, el pla elaborat per a poder solucionar-los consistia bàsicament en un conjunt de sobreestimacions en quant a hores en algunes de les tasques del diagrama de *Gantt*.

Per tant, aquests costos ja han estat inclosos prèviament a la secció del cost de cada activitat, i el que es farà es exposar quins són aquests costos, però no sumarlos al total (ja que com hem dit, ja s'ha sumat anteriorment). Tot això es farà suposant que la possibilitat d'aparició d'aquests imprevistos és d'un 30%, i serà per aquest factor per el que multiplicarem les partides d'imprevistos.

Finalment, veiem que el pressupost total considerat pel disseny i desenvolupament sencer del projecte és de 10609 euros.

## 3.2. Control de gestió

Finalitzem la secció considerant els casos en que serà més possible que apareguin desviacions en el pressupost i aplicar així els indicadors més adequats al llarg del desenvolupament.

Per la naturalesa del treball classificat com a projecte d'investigació, s'ha considerat que la major part de les possibles desviacions seran a les activitats associades al diagrama de *Gantt*. Pel que fa a això, s'utilitzarà com a indicador de desviació d'activitats la següent fórmula: (cost estimat – cost real) * consum real hores de treball.

Altres indicadors que utilitzarem seran per estimar el cost en les possibles desviacions d'un recurs de *hardware* ((cost estimat – consum real) * cost real) o per calcular la desviació total de recursos (cost estimat total – cost real total).

D'aquesta manera, es podrà avaluar de manera dinàmica i parella a la elaboració del treball, quines i com són les desviacions sobre el pressupost.

## 3.3. Lleis i regulacions

És important contemplar l'existència de lleis i normatives rellevants que afecten al projecte i que per tant, s'han de considerar per tal d'assegurar que es compleixen.

Avui en dia el tractament d'informació està molt regulat, i ho està encara més quan aquest tipus d'informació és de caràcter sensible com ho és l'informe i historial clínic d'una persona, que son les dades que tractem dins d'aquest projecte.

Per tant, és important seguir estrictament tots els protocols associats al tractament d'aquestes dades. Per començar, remarcar que les dades que se'ns mostren, no aporten informació que permeti identificar al pacient, buscant així mantenir l'anonimat de la persona a la qual van associades aquelles dades.

D'altra banda, tot tractament que es faci d'aquestes dades, s'ha de fer directament des del clúster del rdlab del Departament de Ciències de Computació de la UPC. Tot usuari que necessiti tractar-les, ho podrà fer únicament des d'aquí, i prèviament haurà signat una sèrie de documents on es fa responsable d'un correcte ús d'aquestes dades, assegurant que no es faran públiques fora d'aquest clúster.

Aquesta és la restricció legal que tenen, i que per tant ens hem d'assegurar que es compleix de manera estricta, aconseguint així que la totalitat de les dades siguin tractades amb seguretat i confidencialitat.

# 4. Informe de sostenibilitat

## 4.1. Impacte ambiental

Actualment vivim un moment en el que estem començant a patir les conseqüències de tot el desenvolupament industrial i tecnològic que hem tingut durant les darreres dècades. Per tant, ara més que mai, és molt important ser conscients d'això i tractar de minimitzar aquesta petjada humana que deixem al nostre planeta.

Per aquest motiu, a continuació analitzarem l'impacte ambiental que té el nostre projecte. Aquest anàlisi el farem tant per aquest apartats com per als següents, a nivell de posta en marxa, vida útil i riscos.

El primer que farem és quantificar en kWh quin és l'impacte ambiental de la realització del projecte. Per a fer aquest càlcul, el que considerarem seran dos punts principals:

- **Ús del portàtil personal**: aquest ús es limita a les primeres implementacions i tests. Com no és un portàtil de prestacions molt elevades, analitzant les característiques s'ha estimat un consum de 0.4 kWh.
- **Ús del HPC (*High Performance Computing*) del rdlab**: aquest és la part que major impacte en el consum té. En ser un computador d'alt rendiment, s'estima un cost de 1.5kWh pel que fa a les prestacions que ens esta oferint pel nostre projecte.

Per tant, quantifiquem l'impacte ambiental inicial del projecte en 1.9kWh.

En base a aquest anàlisis, considerem que l'impacte és força elevat i que s'ha de tractar d'establir alguna mesura per tal de reduir-lo. Amb aquest objectiu en ment, el que s'ha fet és tractar de fer les màximes proves possibles dels nostres programes al portàtil personal, buscant reduir al màxim el nombre d'execucions a fer en el HPC.

La implementació d'aquesta mesura s'ha estimat que redueix en un 30% el consum pertanyent al HPC, restant aquest 0.875kWh, i obtenint així un impacte ambiental definitiu del projecte de 1.275kWh.

Difícilment, en cas que s'iniciés el projecte de nou, seria reduïble, ja que els consums depenen directament de la quantitat d'hores de feina que s'hi dedica, i com les dades sobre les que treballem es troben exclusivament al clúster del HPC, queda molt minvada la capacitat de reduir aquest consum.

Pel que fa a la vida útil del projecte, és important comentar que com a projecte d'investigació, el nostre producte encara no s'utilitza de manera oficial en els sistemes als que va destinats. Tampoc considerem que es millorin altres alternatives actual, ja que totes elles requereixen de centres de computació d'alt rendiment per a poder executar els seus programes.

Un cop dissenyats, implementats i entrenats els nostres models, el cost ambiental durant la seva vida útil es veurà molt reduït. De fet, si en fem un anàlisi global, l'impacte serà positiu ja que s'estaran automatitzant tasques que avui en dia es triguen molt més a fer i conseqüentment consumeixen més recursos.

L'escenari de major consum que ens podem plantejar, on la petjada econòmica del projecte seria la major, és on tenim uns sistemes sanitaris que apliquen al complet els nostres sistemes. No obstant, en aquest cas també estaríem obtenint un impacte positiu amb aquesta automatització de les tasques.

## 4.2. Impacte econòmic

La quantificació econòmica del projecte s'ha fet prèviament a l'apartat *3.1.1 Pressupost*, i al llarg de la seva elaboració ha estat difícil trobar mesures que redueixin aquest impacte, ja que les activitats i recursos que defineixen el pressupost són obligatòries.

Per a definir aquest cost econòmic del projecte tractant que sigui el mínim, el que s'ha fet és destinar bastant esforç a fer una bona planificació i distribució de les tasques, buscant no haver de repetir feina i que les hores que es quantifiquen siguin en la gran majoria útils.

També s'intenta reduir el cost en desplaçaments i recursos humans tractant de resoldre els dubtes i comentaris via correu electrònic, i limitar les reunions presencials a situacions estrictament necessàries.

Amb aquestes dues mesures que s'han pres des de l'inici s'espera que s'hagi reduït l'impacte econòmic en un 5-10%.

És interessant comentar que sí que ens hem ajustat al pressupost inicial, ja que sí que han aparegut alguns obstacles dels que teníem prevists inicialment i per tant hem hagut de destinar les hores contemplades com a sobreestimacions a resoldre'ls.

Com ja s'ha comentat amb anterioritat, l'elaboració d'aquest producte té un cost tant ambiental com econòmic molt més gran en la seva posada en marxa que no pas al llarg de la seva vida útil.

Els costos relacionats amb la seva vida útil seran en gran part els ajustos o actualitzacions que s'hagin d'aplicar al model a mesura que anem observant aquestes possibles modificacions durant la seva vida útil.

No obstant, com això es tracta d'un projecte d'investigació, és encara difícil poder quantificar quin serà l'impacte econòmic del producte final quan aquest estigui en funcionament.

De tota manera, s'espera que aquest impacte econòmic sigui a la llarga positiu, ja que es s'estaran substituint tasques per a les que avui en dia s'utilitzen recursos humans, i que en fer-se de manera manual, tenen en proporció un cost més elevat als costos que podrà tenir el producte un cop aquestes tasques es puguin desenvolupar de manera automàtica.

## 4.3. Impacte social

Considero que en relació a la sostenibilitat, l'impacte social és el punt fort d'aquest projecte ja que té un gran nombre d'implicacions positives per a la majoria de la societat.

En primer lloc, comentar que personalment, la realització d'aquest treball m'ha fet ser conscient de que realment el coneixement d'un Enginyer Informàtic pot tenir uns impactes positius en aspectes socials molt més rellevants del que la gent de fora percep del nostre sector.

És cert que els camps que produeixen majors beneficis econòmics no acostumen a ser els que aporten majors beneficis socials, i tot enginyer hauria de poder plantejar-se això i decidir on és més important dedicar el seu coneixement i esforç.

Com ja s'ha comentat a la introducció, el sector sanitari és un sector crític pel que fa al benestar social. I més enllà de la ideologia de cadascú, contribuir a tenir un sistema sanitari àgil, accessible i de qualitat, ha de ser una responsabilitat moral de tota persona.

Aquest projecte, mitjançant l'automatització i millora d'unes tasques que fins ara es realitzen de manera manual, busca poder desplaçar aquest recursos humans que són de gran importància a altres punts del sector clínic, contribuït a reforçar-lo, millorant-ne la rapidesa i robustesa.

A més, la recol·lecció de dades que el nostre sistema és capaç de fer, pot aportar noves informacions a les branques d'investigació clínica, millorant així les dades de que disposen per a desenvolupar els seus projectes.

És, per tant, una necessitat real ja que tot i que tenim un sistema sanitari potent en relació al que podem veure en mols altres països, sempre és millorable i és en el que ens hem de centrar.

Pel que fa a impactes negatius, no es considera cap cas en que aquest producte pugui afectar de manera negativa a algun sector o grup social. No obstant, és important que tinguem en compte que s'estan tractant dades crítiques com són la salut d'una persona, i que conseqüentment, s'hauran de seguir uns estrictes protocols per evitar els riscos associats a la fuga d'aquestes dades.

Finalment, en base als estudis de sostenibilitat que s'han fet a cadascuna de les tres branques (ambiental, econòmica i social) podem concloure que el balanç és positiu.

Hem vist que l'impacte ambiental és considerable, tot i que és pràcticament inherent a tot projecte tecnològic. També hem vist que l'impacte econòmic existeix tot i tampoc ser desmesurat, i que encara és difícil quantificar-lo per no estar parlant encara d'un producte comercial. Però sobretot el que destaquem és l'aspecte social, on el projecte hi aporta moltes millores i és el que fa que ens decantem per aquest balanç positiu.

No obstant sempre hi ha coses a millorar, i el que hauria de ser el focus (tant d'aquest projecte com la resta dins d'aquest sector) és tractar de reduir al màxim possible la petjada ambiental, tractant de contribuir a frenar totes les conseqüències del canvi climàtic.

# 5. Descripció de les dades

## 5.1. Llenguatge utilitzat

En les tècniques d'aprenentatge supervisat, les dades que s'utilitzen per entrenar i validar el model són de vital importància en els resultats finals. Per tant, l'objectiu d'aquest capítol serà descriure i analitzar el corpus de dades sobre el que treballarem per tal d'entendre'l millor i poden analitzar amb més detall el comportament dels nostres models.

Els documents sobre els que treballarem són informes clínics i historial mèdic, escrits tant en castellà com en català pels metges en el moment de les consultes. Com a tal, el llenguatge mèdic ja té força peculiaritats que el diferencien del llenguatge habitual. Però a més, hem de considerar que en una consulta, el metge escriu a gran velocitat mentre paral·lelament atén al pacient, cosa que afegeix més peculiaritats a aquests documents.

Per tant, en primer lloc ens centrarem en caracteritzar el llenguatge que s'utilitza en el conjunt de dades sobre el que treballarem. Per a fer això, llistarem aquelles característiques que diferencien aquest llenguatge del usat habitualment:

- **Paraules tècniques del domini clínic**: Com és lògic, ens trobarem que moltes de les paraules són tècniques, pròpies de la ontologia mèdica. Exemples d'això poden ser fàrmacs com *Hidrocortisona*, o diagnòstics com *Dispepsia*.
- **Acrònims**: L'ús d'acrònims és molt freqüent. Exemples dels més usats poden ser *RX* (Radiografia), *HTA* (Hipertensió arterial) o *DM* (Diabetis Mellitus).
- **Abreviacions**: La necessitat d'escriure amb gran rapidesa deriva en l'ús de moltes abreviacions com *esq.* per voler dir *esquerra*.
- **Faltes d'ortografia**: Aquesta mateixa rapidesa a l'hora d'escriure causa també que veiem moltes faltes d'ortografia i paraules mal escrites com podria ser *broqnuitis* (bronquitis).
- **Indicadors numèrics i dates**: És freqüent l'ús d'indicadors numèrics per exemple per especificar les quantitats d'alguns medicaments i dates fent referencia a altres visites, operacions, inicis de símptomes i situacions d'aquest estil.

Com podem veure, són múltiples les característiques d'aquest textos i serà important desenvolupar les nostres implementacions pensant en això si volem obtenir uns bons resultats.

## 5.2. Anàlisi estadístic

Seguim analitzant d'una manera més estadística el corpus de dades sobre el que treballarem. Com hem comentat anteriorment, de documents o informes en tenim tant en català com en castellà. A continuació llistem quina quantitat de documents tenim en total i separadament per cada idioma (podem veure els percentatges a la Figura 5.1).

- **Total**: 31767 documents
- **Castellà**: 12829 documents
- **Català**: 18938 documents

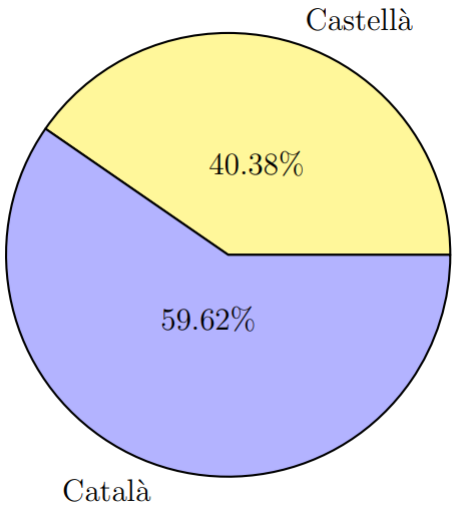


Figura 5.1: *Percentatges dels documents en castellà i català*

Veiem que el nombre de documents en català es significativament més elevat, fet que podrà ser important de considerar de cara a comparar els resultats dels nostres models més endavant.

Seguim amb l'anàlisi de les entitats que apareixen etiquetades en funció de la categoria en que s'hagin classificat. Mostrem el nombre d'aparicions (en percentatges, Figura 5.2):

- **Signe / Símptoma**: 23053 entitats
- **Fàrmac**: 17810 entitats
- **Diagnòstic**: 14940 entitats
- **Part del cos**: 14488 entitats

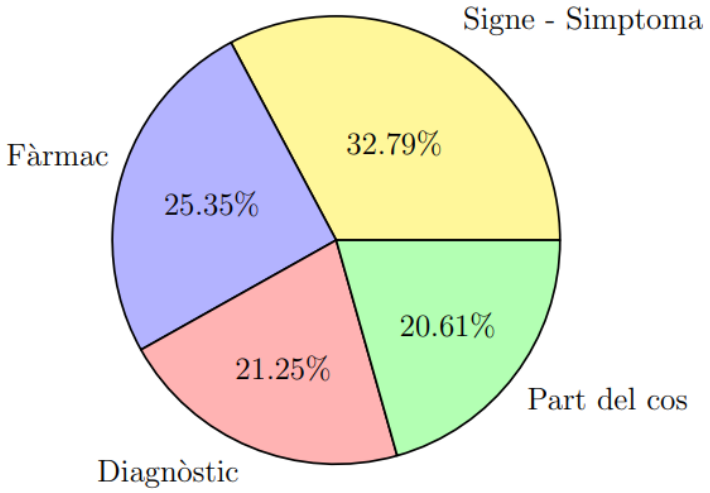


Figura 5.2: *Percentatges en funció del tipus d'etiquetatge.*

Observem que la distribució és més o menys equitativa, destacant la quantitat d'entitats etiquetades com a Signe/Símptoma per sobre de la resta.

Continuant amb l'anàlisi, hem vist als exemples de la part inicial que una entitat pot estar composta d'una única paraula (e.g. hepatitis) o de dos o més paraules (e.g. càncer pulmonar). Per tant, serà també important veure, de cara al posterior disseny i comportament dels nostres models, com es distribueixen les entitats segons el nombre de paraules en funció a la categoria en què estiguin etiquetades.

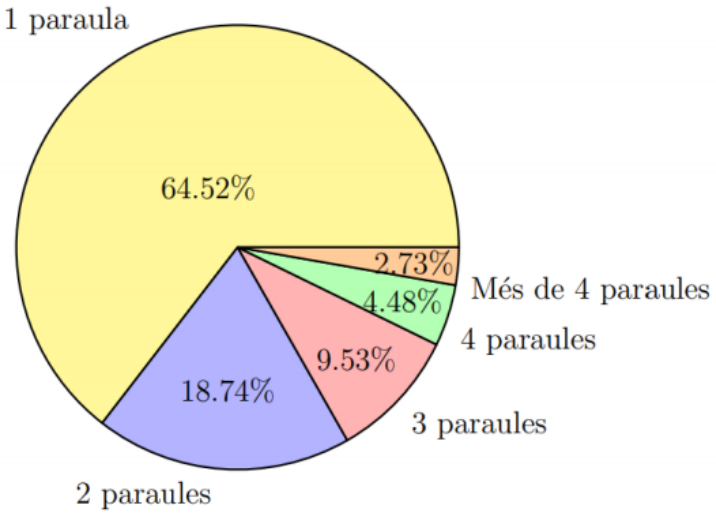


Figura 5.3: *Percentatges en funció del nombre de paraules de cada entitat etiquetada com a Signe/Símptoma.*

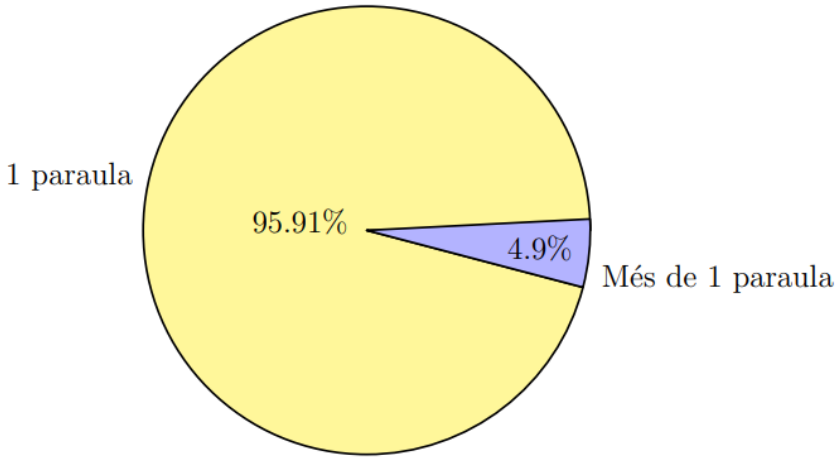


Figura 5.4: *Percentatges en funció del nombre de paraules de cada entitat etiquetada com a Fàrmac.*

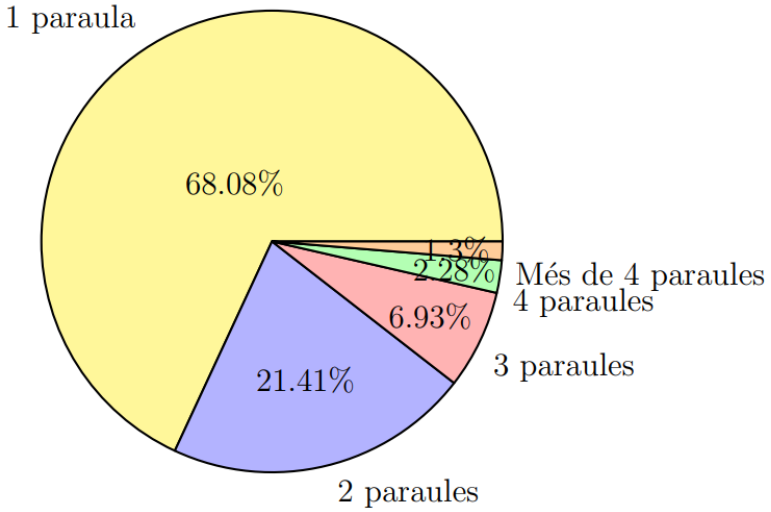


Figura 5.5: *Percentatges en funció del nombre de paraules de cada entitat etiquetada com a Diagnòstic.*

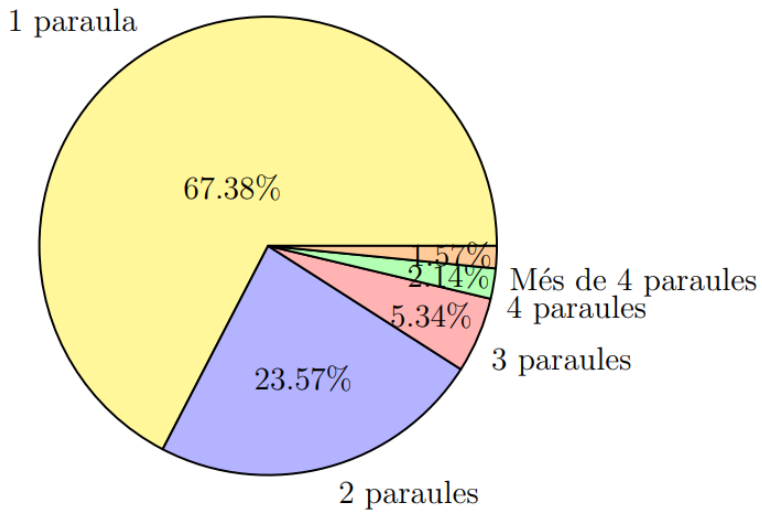


Figura 5.6: *Percentatges en funció del nombre de paraules de cada entitat etiquetada com a Part del Cos.*

Com hem pogut veure, les entitats classificades com a Signe/Símptoma (Figura 5.3), Diagnòstic (Figura 5.5) i Part del Cos (Figura 5.6) segueixen una distribució en funció del nombre de paraules per entitat força similar. En canvi, les entitats classificades com a Fàrmac (Figura 5.4) són pràcticament totes compostes d'una única paraula, fet que ens permetrà explicar alguns comportaments determinats més endavant.

# 6. Metodologia

## 6.1. Introducció a la metodologia

Ja hem analitzat en profunditat les dades sobre les que treballarem, per tant a continuació exposarem la manera en la que tractarem de resoldre el problema al que ens enfrontem.

L'elecció de les tècniques que escollirem per implementar els nostres mètodes són de molta importància, ja que volem combinar la obtenció de bons resultats juntament amb la utilització de tècniques innovadores que contribueixin als avenços en aquest camp d'investigació.

El primer mètode es basarà en la utilització d'un model probabilístic[1] com son els CRFs(*Conditional Random Fields*).

Tot i que en la definició de que són els CRFs i de com els utilitzarem ja ens centrarem més endavant, el que esperem de la utilització d'aquesta tècnica (força validada per altres projectes dins d'aquest camp) és la obtenció d'uns bons resultats, tractant d'implementar-la de la manera més adequada per les característiques del nostre projecte.

De cara a l'elecció de la segona tècnica, la idea era utilitzar alguna alternativa de la branca del *Deep Learning*[2] que ja s'hagi aplicat amb bons resultats a diferents sectors del camp del NER i que, per tant, sigui interessant d'aplicar sobre el nostre problema per a analitzar els resultats que produeixi.

Per aquest segon mètode, s'implementarà una xarxa neuronal artificial basada en una arquitectura Bi-LSTM-CRF(*Bidirectional - Long Short-Term Memory – Conditional Random Field*) que ja definirem més endavant en aquest capítol. Amb això, el que es pretén es suplir algunes de les carències que tindrà el nostre primer model i veure si així podem millorar els resultats obtinguts en aquest.

[1]Model matemàtic que es basa en els supòsits estadístics que es fan sobre la generació de dades mostrals.

[2]Subconjunt de tècniques del *Machine Learning* que generalment utilitzen xarxes neuronals artificials amb arquitectures que passen la informació a través de diferents nivells jeràrquics.

Com s'ha comentat, aquest dos mètodes els definirem amb més precisió en futures seccions dins d'aquest capítol. Però, en primer lloc, ens hem de centrar en una part prèvia a la implementació d'aquests mètodes i necessària per a tots dos, com serà el pre-processat, tokenització i etiquetatge de les dades. Aquesta part la tractarem a la següent secció.

## 6.2. Processament i tractament de les dades

Abans de poder començar a implementar els nostres mètodes, hem de tractar i processar les dades per a passar-les a un format amb el que els nostres models puguin treballar. Aquest processament és molt important, ja que les diferents decisions que anem prenent en aquesta part tindran un impacte important en els resultats finals.

Aquest tractament i transformació de les dades el dividirem en les següents tres fases:

- **Preprocessat**
- **Tokenització**
- **Etiquetatge**

A continuació, s'exposarà de manera detallada cadascuna de les fases.

### 6.2.1. Preprocessat

Aquesta fase inicial consisteix en fer un tractament de les dades de manera que posteriorment els nostres models puguin treballar millor sobre aquestes.

Generalment, en aquest procés el que es fa és eliminar algunes característiques del text de manera que es faciliti el treball posterior, reduint ambigüitats i possibles confusions i, en definitiva, deixant el text de manera que el nostre model tingui major rendiment.

No obstant, en el nostre cas en que estem desenvolupant dos tècniques diferents, no podem aplicar el mateix tipus de preprocessat per les dues ja que el seu comportament és diferent i no seria vàlid generalitzar-lo per als dos models.

En primer lloc, en els dos casos el que farem és extreure els accents de les paraules. Aquesta és una bona pràctica en base a la tipologia dels textos sobre els que estem treballant, ja que com hem comentat anteriorment, les equivocacions en la ortografia són freqüents, i els accents, que no sempre apareixeran correctament indicats, tendiran a causar confusions.

Un preprocessat que s'ha plantejat aplicar és la transformació de totes les lletres a minúscules. No obstant aquesta és una transformació que té més impacte i s'ha d'analitzar si és correcta d'aplicar en cada mètode.

En el cas dels CRF, com veurem a la secció *6.3 CRFs (Conditional Random Fields)*, aquests es basen en determinar i utilitzar diferents característiques del text per a realitzar les prediccions, i per tant, estaríem causant una pèrdua d'informació en aquest model si transforméssim totes les lletres a minúscules, ja que les lletres majúscules ens podrien estar indicant noms propis o paraules rellevants, i conseqüentment estaríem perdent tota aquesta informació.

Aquesta transformació, en canvi, si que té sentit en el nostre model de xarxa neuronal, ja que com definirem amb més exactitud a la secció *6.4 Xarxes neuronals Bi-LSTM-CRF*, utilitzarem uns *word embeddings* pre-entrenats, i per tant necessitarem tenir les paraules en minúscules per a poder associar-les a les paraules que apareixen als *embeddings*. De fet, s'han fet diferents proves on s'ha confirmat que aquesta transformació a minúscules millora el comportament.

Aquí finalitza el preprocessat que es fa sobre el text. Tractar d'estendre'l més pot derivar en massa pèrdua de característiques del text que ens redueixin la capacitat de predicció dels nostres models. Per exemple, no voldríem l'eliminació de caràcters com '%' o altres signes, ja que en els informes clínics generalment s'utilitzen propers a fàrmacs o altres entitats a detectar, i no volem perdre aquesta informació.

A més, a la secció de preprocessat no s'han tractat els signes de puntuació ja que com veurem a continuació, aquest tractament pertany a la fase de Tokenització.

### 6.2.2. Tokenització

El procés de tokenització és molt important en tot projecte de NLP, aquest consisteix en separar un text en diferents unitats o tokens. Aquesta separació es pot fer a diferents nivells, un exemple d'això seria tokenitzar un document sencer separant-lo en les frases que el formen, on cada frase seria un token.

En el nostre cas, però, el que ens interessa és tokenitzar a nivell de paraula. Més concretament, ens basarem en el llibre *Introduction to Information Retrieval*[7] per definir un token com una instancia d'una seqüencia de caràcters en un document particular que es troben agrupats formant una unitat semàntica amb significat.

Si bé inicialment pot semblar tant senzill com separar segons espais, salts de línia o signes de puntuació, realment no ho és tant. El llenguatge humà està ple d'ambigüitats i ens podem trobar moltes situacions especials on el tokenitzador hagi

d'aplicar un tractament especial, per exemple quan ens trobem apòstrofs (e.g. *l'edema*), barres per fer referencia a dosis de medicaments (500mg/8h), abreviatures o signes de puntuació en números.

Davant tot això, existeixen diferents alternatives per a dur a terme aquesta tokenització. Algunes d'elles es basen en la utilització d'expressions regulars[3] .D'altres es basen en entrenar un model per a que sigui capaç de realitzar automàticament aquesta separació en tokens.

En el nostre cas, utilitzarem una eina ja molt consolidada a la comunitat com és el NLTK (*Natural Language Toolkit*)[4], una llibreria dissenyada per al processament del llenguatge natural i que disposa de moltes prestacions, entre elles, un tokenitzador a nivell de paraules, que és el que aplicarem.

A banda d'aquest tokenitzador, serà necessari fer un postprocessat per acabar d'ajustar els tokens resultants a la manera en que ens indiquen les entitats al corpus de dades.

Un cop fet això, ja disposem de la seqüencia de tokens que conforma el document clínic que estem tractant, per tant a continuació el que faltarà és associar a cada token l'etiqueta que corresponent.

### 6.2.3. Etiquetatge

L'objectiu d'aquesta darrera fase corresponent al processament i tractament de les dades, és associar a cada token derivat de la fase de tokenització, l'etiqueta corresponent.

Això el que significa és que indicarem quins dels tokens corresponen a entitats que pertanyen a alguna de les categories sobre les que estem classificant i quins no ho fan. I per aquells que sí que ho fan, haurem d'indicar a quina categoria pertanyen.

Dels diferents esquemes que existeixen per a fer aquest etiquetatge (*labeling*) en el nostre cas utilitzarem dues opcions: BIO i BIOS. Aquests esquemes el que ens permeten es identificar aquells tokens que no corresponen a cap categoria, i identificar aquells que si que ho fan, permetent indicar que més d'un token pertanyen a una mateixa entitat (recordem, cas en que una entitat està composta de diferents paraules).

[3]Seqüències de caràcters que conformen un patró de cerca.
[4]`https://www.nltk.org/api/nltk.html`

Comencem explicat l'esquema d'etiquetatge BIO que el veiem representat en un exemple a la Figura 6.1. El que fem és identificar que un token és el principi d'una entitat (o que ell sol conforma aquella entitat) fent que l'etiqueta comenci amb "B-"(*begining*), i en canvi, quan un token pertany a alguna de les paraules que no son la primera de l'entitat, farem que l'etiqueta vagi precedida de "I-"(*inner*).

D'altra banda, quan un token no pertany a cap entitat, el que es fa es classificar-lo amb una etiqueta especial que en aquest cas és "O".

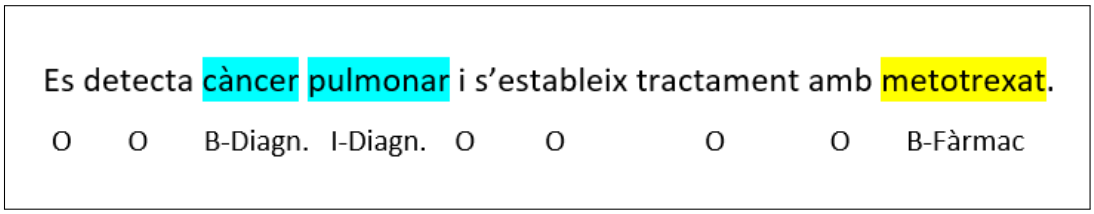


Figura 6.1: *Exemple d'etiquetatge amb esquema BIO.*

Ens centrem ara en l'esquema d'etiquetatge BIOS, que és una extensió de l'anterior, on el que es fa es etiquetar de manera diferent aquells tokens que conformen entitats formades únicament per una paraula. En aquest cas, l'etiqueta d'aquests tokens va precedida per "S-"(single). Podem veure un exemple d'aquest tipus d'etiquetatge a la Figura 6.2.

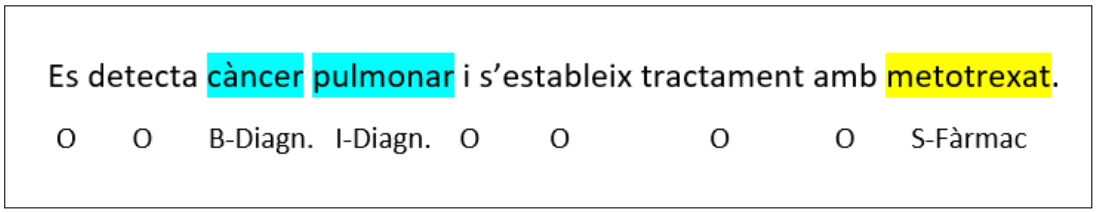


Figura 6.2: *Exemple d'etiquetatge amb esquema BIOS.*

Usualment el comportament de l'esquema BIOS aporta millors resultats quan les entitats que estem etiquetant estan formades en la gran majoria per un únic token, com és el cas en el que estem treballant. No obstant serà interessant aplicar els dos esquemes per a veure quins resultats ens aporta cada un.

## 6.3. Utilització de les dades

A la secció anterior hem fet tot el tractament necessari a les dades per a poder començar a treballar amb elles. Per tant, el que ara determinarem és com utilitzarem aquestes dades per aconseguir que els nostres resultats siguin rigorosament correctes, i que per tant, arribem a conclusions que siguin vàlides.

En primer lloc, exposarem els tres grups en els que dividirem tot el corpus sencer de dades:

- ***Training set***: aquest conjunt de dades serà el que s'utilitzi per a l'entrenament dels nostres models (e.g. pesos de les connexions entre les neurones en una xarxa neuronal artificial).
- ***Validation set***: set de dades que s'utilitza per la validació i ajustament dels paràmetres en l'entrenament del model evitant que aquest estigui esbiaixat.
- ***Test set***: conjunt de dades finals sobre el que s'aplica el model per a avaluar els resultats. No ha tingut res a veure en el procés d'entrenament de manera que evitem que el model estigui esbiaixat.

Per tant, veiem que l'objectiu clar d'aquesta divisió és aconseguir uns resultats que no estiguin esbiaixats i que per tant siguin suficientment representatius com per a permetre'ns extreure unes conclusions vàlides.

Els percentatges que s'utilitzaran per establir les mides de cada una de les particions de les dades, s'exposaran més endavant a la secció corresponent a l'experimentació.

A banda d'això, ja hem vist anteriorment que el corpus sobre el que treballem no és especialment extens, i això pot fer que les seccions de training, validation i test que escollim en el moment inicial estiguin condicionant les característiques dels nostres models.

Per tal d'evitar aquesta situació, aplicarem *5-Fold Cross Validation*. Aquesta tècnica el que ens permetrà és entrenar i avaluar el nostre model 5 cops, cadascun d'ell utilitzant un conjunt de dades d'entrenament i d'avaluació diferents.

Amb això el que aconseguim és que el model no estigui esbiaixat o condicionat pel conjunt de dades de test i avaluació considerats de manera inicial, ja que el que fem es provar diferents seleccions d'aquestes per a obtenir finalment uns resultats més genèrics i per tant més concloents.

## 6.4. CRFs (*Conditional Random Fields*)

Els CRFs són un tipus de model probabilístic que per la seva capacitat de modelar dades seqüencials, s'utilitzen molt en diferents camps del NLP com pot ser el l'etiquetatge POS (Part Of Speech)[5] o el NER, que és la tasca que estem desenvolupant nosaltres.

A continuació entrarem més a fons en com es classifiquen i defineixen els CRFs, però en primer lloc farem unes definicions que tractarem d'ara en endavant:

- *X*, serà qualsevol variable aleatòria de les dades d'entrada que nosaltres volem etiquetar.
- *Y*, serà qualsevol variable aleatòria sobre les seqüències d'etiquetes.

Dins el camp del *Machine Learning*, podem classificar els diferents algoritmes o tècniques com a generatius o discriminatoris.

Els de tipus generatiu el que fan és proporcionar un model de com es generen les dades. És a dir, aprenen una distribució de probabilitat conjunta[6] P(X,Y), i a partir del que han aprés, poden classificar les següents observacions de variables que hagin de tractar.

En canvi, els classificadors discriminatoris, com els CRFs, el que fan és aprendre una distribució de probabilitat condicional[7] P(X/Y), tractant de preveure quina serà la variable Y que s'assignarà en funció de la Y.

Un cop entès això, veiem la definició formal de CRF[8]:

> **Definició:** *Donat el graf G = (V,E) tal que* $Y = \boldsymbol{Y}(\boldsymbol{Y}_v)_{v \in V}$*, amb* $\boldsymbol{Y}$ *indexat pels vertex de G. Llavors* $(\boldsymbol{X}, \boldsymbol{Y})$ *és un conditional random field en el cas que, condicionades per* $\boldsymbol{X}$*, les variables aleatòries* $\boldsymbol{Y}_v$ *segueixen la propietat de Markov[9] respecte al graf:* $(\boldsymbol{Y}_v | \boldsymbol{Y}, \boldsymbol{Y}_w, w \neq v) = p(\boldsymbol{Y}_v | \boldsymbol{X}, \boldsymbol{Y}_w, w \sim v)$ *on* $w \sim v$ *significa que w i v son veïns a G.*

[5]Procés d'assignar a cada paraula d'un text la seva categoria gramatical.

[6]Probabilitat de la intersecció dels events X i Y, és a dir, probabilitat de que succeeixin de manera simultània.

[7]Probabilitat de que succeeixi un event X, sabent que s'ha produït un event Y.

Aquesta definició, aplicada sobre el nostre problema de modelatge de seqüencies, resulta en un graf en forma de cadena com el que podem veure a la Figura 6.3:

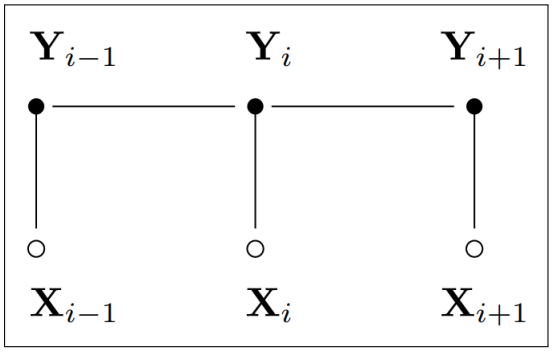


Figura 6.3: *Exemple d'un CRF amb estructura de cadena per seqüencies. Font: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[8].*

El nostre objectiu, per tant, serà modelar aquest graf determinant el conjunt de *features* amb el que aportarem la informació per caracteritzar cada token (X) i que posteriorment el model sigui capaç de predir l'etiqueta associada (Y).

## 6.5. Xarxes neuronals Bi-LSTM-CRF

### 6.5.1. *Word embedding*

Abans d'explicar l'arquitectura que utilitzarem per aquesta segona tècnica basada en xarxes neuronals, introduirem el concepte de *word embedding* i veurem com i perquè ens és d'interès per al nostre projecte.

Els word embeddings són un conjunt de tècniques de modelatge del llenguatge[8] que es basen en fer un mapeig de les paraules o frases a un espai vectorial utilitzant nombres reals.

Aquest conjunt de tècniques ja s'aplicava feia temps al camp estadístic del modelatge del llenguatge[10], però no és fins als darrers avenços en els mètodes basats en xarxes neuronals i el desenvolupament de tècniques com el Word2Vec (que tractarem més endavant en aquesta secció), que no havien adquirit aquest elevat grau de popularitat que tenen avui en dia.

[8]Distribució de probabilitat sobre una seqüència de paraules. Donada una seqüència de longitud *n*, se li assigna una probabilitat de *P(w1,...,wn)* a tota la seqüència.

Aquests word embeddings, en funció de com es construeixen i de quin corpus de dades utilitzen en el moment de ser entrenats, aporten un coneixement determinat. Això vol dir que si han estat entrenats utilitzant un corpus genèric de la llengua anglesa, funcionaran bé per documents de caire genèric escrits en anglès, però no ho faran tant per documents d'una temàtica més específica o en un altre llenguatge.

Per tant, si la nostra idea és utilitzar un word embedding ja construït (*pre-trained*) haurem d'utilitzar un el més proper possible a la tipologia de documents que tractem en aquest treball.

En aquest context, tenim al nostre abast el projecte *Medical Word Embeddings for Spanish: Development and Evaluation*[11] on es construeixen uns word embeddings basats en la ontologia mèdica per la llengua castellana, i que per tant, seran els que s'utilitzaran per desenvolupar el nostre sistema.

En aquest projecte que hem mencionat, per a construir el model es basen en un corpus obtingut a partir de dos fonts de dades:

- SciELO (*Scientific Electronic Library Online*)[9], on existeixen un seguit d'articles mèdics escrits en llengua castellana.
- Wikipedia[10], d'on agafen un subconjunt d'articles relacionats amb el domini clínic.

I a més, utilitzen dos mètodes diferents per generar els embeddings: Word2Vec i Fasttext. Tot i que no ens endinsarem en detall a explicar com genera cada mètode els embeddings, sí que serà interessant comentar les principals diferencies entre ells de cara a l'aplicació en el nostre projecte.

La tècnica de Word2Vec va ser desenvolupada i patentada per un grup d'investigadors de Google al 2013[12]. Aquesta tècnica es basa en tractar cada paraula com a una entitat atòmica, i generar per a cada una d'aquestes un vector. En aquest sentit, Word2Vec funciona molt bé quan la majoria de paraules dels documents que es tractaran, les podrem trobar als embeddings escrites de la mateixa manera.

Aquest, però, no serà el nostre cas, ja que com podem recordar, estem tractant uns informes clínics on les faltes i equivocacions en l'ortografia seran molt freqüents, fet que farà més difícil la tasca d'associar la paraula al word embedding.

En aquest context, situem l'altre tècnica amb la que generen els embeddings, el Fasttext. Aquesta tècnica és fonamentalment una extensió de l'anterior, que va ser proposada per Facebook al 2016[13].

[9] `scielo.org`
[10] `es.wikipedia.org`

La idea bàsica del seu funcionament és, que a diferencia del Word2Vec que tracta cada paraula com una unitat no fragmentable, el Fasttext tracta cada paraula separant-les en diversos n-grames[11]. En el projecte que estem utilitzant com a referència pels word embeddings, s'utilitza n-grames de com a mínim 3 i com a màxim 6 caràcters. Per tant, un a paraula com *"dolor"* es fragmentaria de la següent manera: *"dol", "dolo", "dolor", "olo", "olor", "lor"*.

El que això ens aporta és un millor tractament per aquelles paraules que ja sigui per estar fora del vocabulari que contemplen, o perquè estan malament escrites (que per nosaltres serà un cas especialment freqüent), no aconseguim associar directament a cap paraula del word embedding.

Per tant, les dues tècniques seran utilitzades i avaluades en la capa que formarà els embeddings en l'arquitectura de la nostra xarxa neuronal, que s'exposarà de forma detallada a la següent secció.

### 6.5.2. Arquitectura Bi-LSTM-CRF

La utilització d'aquest tipus d'arquitectura en xarxes neuronals, ha aportat resultats significativament positius en diferents projectes desenvolupant tasques de NER[14][15][16][17].

Per tant, aquesta serà l'arquitectura que implementarem, acompanyada d'una capa inicial que seran els word embeddigns que hem comentat a l'apartat anterior.

Les LSTM (*Long short-term memory*) són un tipus de xarxa neuronal recurrent (RNN)[12] proposades al 1997 i que per la seva capacitat per trobar patrons en llargues seqüències de dades, s'utilitzen molt en camps com el reconeixement d'escriptura manual (*Handwriting recognition*, HWR) o el reconeixement de la parla (*Speech recognition*).

Tot i que no entrarem en la definició detallada de com és una unitat LSTM, el que la diferència de les RNNs comuns, és que les actualitzacions a les captes ocultes es substitueixen però unes cel·les de memòria dissenyades específicament. D'aquesta manera, les LSTM poden tractar i explotar dependències en seqüències de dades molt més llargues, pel que es podria considerar que tenen una capacitat de memòria més elevada.

[11] Un n-grama, és una subseqüència de n elements d'una seqüencia inicial. En el nostre cas, els n-grames serien totes les subseqüències possibles de n caràcters contigus en una seqüència donada.

[12] Tipus de xarxes neuronals on les connexions entre els nodes formen un graf dirigit al llarg d'una seqüència temporal. Això permet que els nodes es puguin retroalimentar, aconseguint temporalitat i permetent que la xarxa tingui memòria.

Les LSTM, preserven la informació que ja ha passat a través d'elles utilitzant l'estat ocult. Aquesta informació, en les LSTM unidireccionals, prové únicament del passat (en el nostre cas, del text previ al punt on ens trobem) ja que les entrades que les cel·les poden veure són del passat. Això ho podem veure representat a la Figura 6.4.

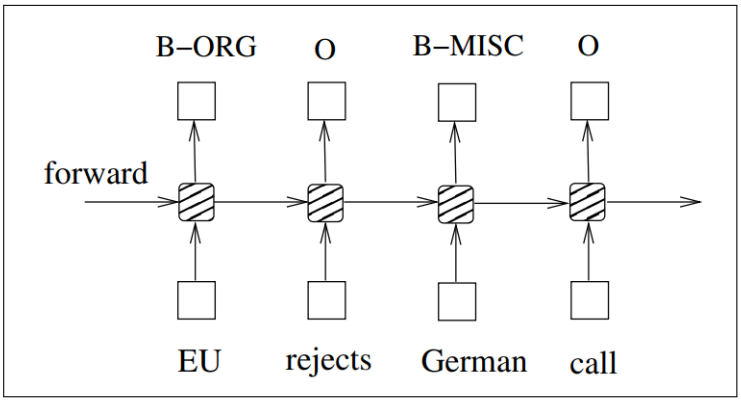


Figura 6.4: *Exemple de model de xarxa neuronal LSTM. Font: Bidirectional LSTM-CRF Models for Sequence Tagging[14].*

Això el que implica és que en el moment de tractar un determinat token, únicament s'estigui utilitzant el context del text previ a l'aparició d'aquest. Però que passa si volem utilitzar tant el context sencer del document on apareix aquest token? En aquest cas entren al joc les Bi-LSTM (*Bidirectional* - LSTM).

Les Bi-LSTM, el que fan és tractar la informació tant d'endavant cap al darrere com al revés, aconseguint així mantenir el context tant del que apareix abans com darrere del punt en el que ens trobem del text, tal i com podem veure a la Figura 6.5.

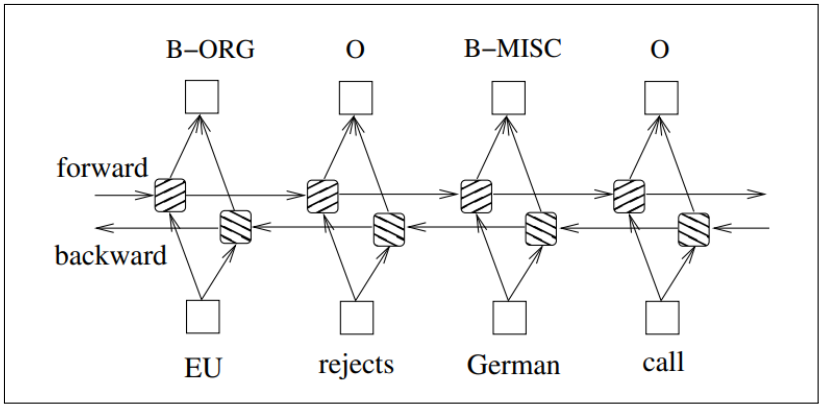


Figura 6.5: *Exemple de model de xarxa neuronal BiLSTM. Font: Bidirectional LSTM-CRF Models for Sequence Tagging[14].*

Un cop tenim això, s'ha demostrat que per tasques d'etiquetatge de seqüències, l'aplicació d'una darrera capa de CRFs, afegeix una sèrie de restriccions que la capa va aprenent duran el procés d'entrenament, i les aplica al moment de determinar l'etiqueta que s'associarà al token que tractem, assegurant-se que aquesta es vàlida.

Per tant, l'arquitectura definitiva del nostre model de xarxes neuronals constarà de:

- **Embedding Layer** (exposada a l'apartat anterior)
- **Bidirectional LSTM**
- **CRFs**

# 7. Experimentació i resultats

## 7.1. Mètode 1: CRFs

En aquesta secció ens centrarem a exposar l'experimentació que s'ha efectual sobre el model basat en CRFs, justificant de manera incremental el conjunt de *features* que s'ha utilitzat, i analitzant els resultats amb aspectes positius i negatius del nostre model.

Aprofitem per comentar que s'han plantejat dos tipus d'avaluacions de resultats: estricta, on es considera una entitat anotada correctament quan s'han etiquetat de manera correcta totes les seves paraules, i parcial, on es consideren aquelles entitats que no s'han etiquetat la totalitat dels seus components com a vàlides (e.g. *càncer pulmonar*, només s'etiqueta com a Diagnòstic *càncer*).

En aquest cas, ens limitarem a aplicar l'avaluació estricta, essent aquesta la més restrictiva i la que ens aporta la cota inferior en quant a puntuacions obtingudes.

### 7.1.1. Features i context

Primerament, en aquest apartat ens centrarem en experimentar sobre quins són els millors conjunts de features a utilitzar i sobre quins elements del context els apliquem.

Recordem que els features són aquelles característiques que recollim sobre un determinat token, i que poden fer tant referència a ell mateix com als elements del seu context. Aquestes característiques són les que el model utilitzarà per entrenar i poder predir quina etiqueta li assignarem a aquell token determinat.

Per a l'experimentació dividirem el conjunt de dades que tenim, deixant un 20% d'aquestes com a conjunt de test per avaluar els resultats de les diferents execucions que realitzem.

Aquest anàlisi el farem separant els documents en castellà i català, de manera que puguem observar els resultats més al detall, podent estudiar per separat les característiques de cada llengua.

Començarem definit el conjunt de features que millor funcionen sobre el token actual, és a dir, el token que estem tractant en un moment determinat i sense tenir en compte els del context.

Això ho farem de manera incremental: anirem incloent aquells que ens aportin millors resultats, i el que no ho fàcil els descartarem.

Per a fer tot aquest anàlisis ens centrarem en la f1-score[1], i deixarem precisió[2] i el *recall*[3] per a mes endavant.

Comencem analitzant la millor manera de guardar el valor del token. Contemplem dos possibilitats, guardar-lo tal ens apareix (`word`) o guardar el token transformat a minúscules (`word.lower`). A la Taula 7.1 veiem que els millors resultats són clarament guardant-lo en minúscula, per tant mantindrem aquest feature.

També observem que els resultats en castellà son significativament millors que en català, una característica que continuarà present a la resta de l'experimentació.

| | **Castellà** | **Català** |
|---|---|---|
| `word` | 0.47203 | 0.43608 |
| `word.lower` | 0.56181 | 0.51322 |

Taula 7.1: *Resultats per els features word i word.lower.*

Tot seguit veurem si ens compensa afegir el prefix i el sufix del token actual. Recollir aquestes característiques pot ser molt interessant de cara a tractar terminologia clínica, que sabem que tenen prefixos i sufixos força freqüents (e.g. *anti-*, *-itis*).

A més d'això, comprovem quins finestres ens pot interessar recollir (nombre de caràcters que seleccionem). L'experimentació de la Taula 7.2 és acumulativa (quan provem finestra 4, estem incloent 2 i 3, i ho fem tant per davant (prefix) com per darrere (sufix)). Com podem veure, els millors resultats s'obtenen recollint fins a 2,3,4 i 5 caràcters.

Plantegem el feature `word.pattern` on el que farem és quedar-nos amb el patró de la paraula, substituint les lletres minúscules per "a", les majúscules per "A", els números per "0ï els signes de puntuació per "-".

[1]Mètode de puntuació que combina tant la precisió com el recall.

[2]Mètode de puntuació que ens indica quants dels elements que hem seleccionat, corresponen a positius reals.

[3]Mètode de puntuació que reflecteix els elements que han sortit positius en relació als que haurien d'haver sortit positius en realitat.

| | Castellà | Català |
|---|---|---|
| **2 caràcters** | 0.57085 | 0.54116 |
| **3 caràcters** | 0.57823 | 0.55825 |
| **4 caràcters** | 0.58013 | 0.56778 |
| **5 caràcters** | 0.58192 | 0.56935 |
| **6 caràcters** | 0.58036 | 0.56563 |

Taula 7.2: *Resultats per els features sobre els prefixos i sufixos.*

Amb això el que pretenem es detectar aquells tokens que tenen una estructura predeterminada (e.g *Ibuprofeno(600mg)* passaria a ser *Aaaaaaaaaa-000-aa*), que són força freqüents en quests documents. A la Taula 7.3 veiem com efectivament tenim una millora dels resultats.

| | Castellà | Català |
|---|---|---|
| `word.pattern` | 0.59513 | 0.57725 |

Taula 7.3: *Resultats del feature word.pattern*

Seguim amb dos features enfocats al tractament dels tokens compostos no només de lletres. En primer lloc tenim `word.removeLetters` on el que tindrem és el token inicial però eliminant tot caràcter de l'alfabet. D'aquesta manera ens quedarem només amb els números i signes. Els resultats de l'aplicació d'aquest feature són positius com podem veure a la Taula 7.4.

D'altra banda, considerarem el feature invers: `word.justLetters`, eliminarà del token tot caràcter que no sigui pertanyent a l'alfabet. En aquest cas, però, no obtenim millors resultats (Taula 7.5) i per tant descartem el seu ús.

| | Castellà | Català |
|---|---|---|
| `word.removeLetters` | 0.60113 | 0.58076 |

Taula 7.4: *Resultats del feature word.removeLetters*

Continuem amb dos features que determinen si la característica que estem contemplant es compleix o no. Primerament tenim `word.hasPunct` (Taula 7.6) que determina si el token que estem tractant té algun signe de puntuació. En el nostre cas, com a signes de puntuació considerem el següent ventall: !"$%'()*+,‘−./ :;<=>?@[]}{|_.

En segon lloc, tenim el feature `word.hasLetNum` (Taula 7.7) on dictem si el token compleix dues condicions que són tenir com a mínim un caràcter de l'alfabet i com a mínim un caràcter numèric.

| | Castellà | Català |
|---|---|---|
| `word.justLetters` | 0.59583 | 0.57976 |

Taula 7.5: *Resultats del feature word.justLetters*

Amb aquests dos features, seguim en la línia de trobar aquelles paraules o simbologia que utilitzen específicament els metges als informes clínics, com és per exemple indicar tants per cents o altres usos més específics de cada usuari.

Com podem observar, tots dos tokens aporten millores i per tant ens interessarà conservar-los per al nostre model.

| | Castellà | Català |
|---|---|---|
| `word.hasPunct` | 0.60779 | 0.59277 |

Taula 7.6: *Resultats del feature word.hasPunct*

| | Castellà | Català |
|---|---|---|
| `word.hasLetNum` | 0.60909 | 0.59282 |

Taula 7.7: *Resultats del feature word.hasLetNum*

Prosseguim tres features de caràcter més genèric: `word.isLower`, `word.isUpper` i `word.isTitle` que defineixen si una paraula està tota en minúscules, tota en majúscules o la primera lletra en majúscules i la resta en minúscules, respectivament.

La idea de fons en la utilització d'aquestes features és fer més èmfasi en la detecció de noms propis (que sovint s'escriuen començant per lletra majúscula) o aprofitar el fet que ens molts casos les paraules claus com els Fàrmacs les escriuen tota en majúscules.

En Taula 7.8 s'observen els resultats corresponents a cada un dels features comentats, però en cap cas s'aconsegueixen millors resultats, per tant no els utilitzarem al nostre model.

| | Castellà | Català |
|---|---|---|
| **word.isLower** | 0.60892 | 0.59235 |
| **word.isUpper** | 0.60902 | 0.59042 |
| **word.isTitle** | 0.60813 | 0.59268 |

Taula 7.8: *Resultats dels features word.isLower, word.isUpper, word.isTitle*

Amb la voluntat de generalitzar el model i de introduir coneixement expert, s'han elaborat quatre llistats on s'exposen els prefixos i sufixos més freqüents per fàrmacs i diagnòstics. Amb això, a part, també es pretén reduir la possible confusió del nostre model en determinar aquestes dues categories.

Per tant, crearem els nous quatre features que tenim a continuació per determinar si el prefix/sufix pertany als més comuns en fàrmacs/diagnòstics:

- `word.topFRMpref`: Prefix pertanyent als prefixos de fàrmacs més comuns.
- `word.topFRMsuff`: Sufix pertanyent als sufixos de fàrmacs més comuns.
- `word.topDGNpref`: Prefix pertanyent als prefixos de diagnòstics més comuns.
- `word.topDGNsuff`: Sufix pertanyent als sufixos de diagnòstics més comuns.

Els resultats de la utilització dels quatre features en conjunt el podem veure a la Taula 7.9. Observem que hi ha una milloria, i per tant els preservem.

| | Castellà | Català |
|---|---|---|
| **Suf/Pref comuns** | 0.61071 | 0.59375 |

Taula 7.9: *Resultats de l'aplicació del conjunt de features per a prefixos/sufixos de fàrmacs/diagnòstics comuns.*

Finalitzarem elaborant dos darrers features que sovint aporten bons resultats i que determinen si una paraula es troba al principi d'un document (BOS, *Begin of Sentence*) o al final (EOS, *End of Sentence*). En aquest cas veiem també una millora dels resultats a la Taula 7.10.

| | Castellà | Català |
|---|---|---|
| **BOS / EOS** | 0.61306 | 0.59640 |

Taula 7.10: *Resultats del feature per determinar BOS / EOS.*

Concloent, hem analitzat de manera incremental quins són els features que ens aporten millors resultats i hem descartat aquells que no ho fan. Per tant, seguidament llistem quin és el conjunt de features que aplicarem sobre el token que estem tractant en aquell determinat moment (sense tenir en compte context encara):

- Token en minúscula.
- Prefixos i sufixos agafant des de 2 caràcters a 5.
- Patró del token (Majúscula → "A", Minúscula → "a", Número → "0", Signe → "-").
- Només els caràcters del token que no són lletres.
- Si el token té signes de puntuació.
- Si el token té lletres i números.
- Si el token pertany els prefixos de fàrmacs més comuns.
- Si el token pertany als sufixos de fàrmacs més comuns.
- Si el token pertany als prefixos de diagnòstics més comuns.
- Si el token pertany els sufixos de diagnòstics més comuns.
- Si el token correspon a l'inici del document.
- Si el token correspon al final del document.

Fins ara hem estat definint el conjunt de features que millor funciona sobre l'element actual que estem tractant en un moment determinat, sense tenir en compte cap altre element del context.

A continuació experimentarem incloent de diverses maneres elements del context per a veure com poden portar millors resultats al nostre model.

En aquest punt, considerarem dues maneres d'introduir característiques del context. En primer lloc, podem introduir aquestes característiques tenint en compte l'ordre dels elements del context, és a dir, tindrem en compte la paraula que es troba exactament una posició a la dreta, quina a dos posicions, i així consecutivament (igual per les paraules situades a l'esquerra).

La segona manera, serà utilitzant el model de BoW (*Bag of Words*)[4] on el que farem es recollir aquelles paraules que es troben a l'esquerra i a la dreta en dos "bosses"diferents, sense tenir en compte en quin ordre apareixen respecte del token actual.

En primer lloc explorarem la opció d'utilitzar l'ordre dels tokens del context. Per fer això utilitzarem diferents finestres (agafant des d'una paraula per davant i per darrere fins a 5), mantenint per a totes elles l'ordre en que es troben en relació al token actual.

Per a cada un dels tokens del context, aplicarem també el mateix conjunt de features que hi apliquem sobre el que token actual.

Els resultats els veiem a la Figura 7.1 on observem que tant pel català com pel castellà, els millors resultats s'obtenen agafant la primera paraula a la dreta i la primera a l'esquerra. Aquests millors resultats són de 0.63149 pel castellà i 0.61529 pel català.

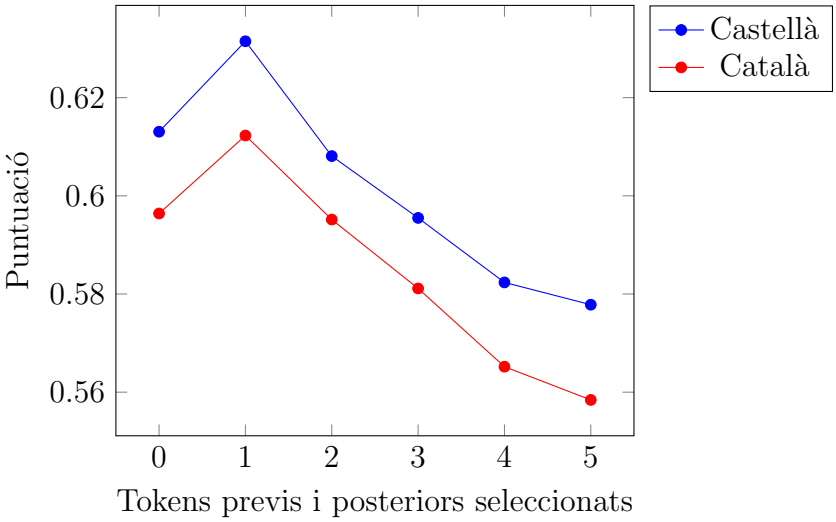


Figura 7.1: *Gràfica dels resultats d'utilitzar tokens del context amb ordre.*

A continuació, realitzarem el mateix tipus d'experimentació però utilitzant en aquest cas la tècnica de BoW. Per aquest experiment, no recollirem els features per les paraules del context, simplement ens quedarem amb elles.

[4]Mètode utilitzat al processament del llenguatge on s'ignora l'ordre de les paraules, simplement es considera si aquesta paraula apareix o no.

Veiem en els resultats de la Figura 7.2 que el millor rendiment s'obté quan agafem una finestra d'una paraula per davant i una per darrere. En aquest cas les millors puntuacions són de 0.63563 pel castellà, i de 0.61444 pel català, millorant els resultats de l'experimentació anterior.

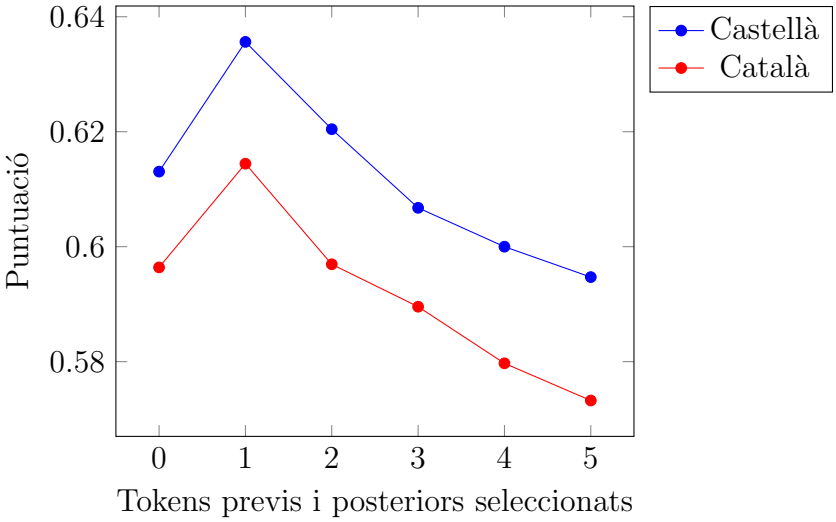


Figura 7.2: *Gràfica dels resultats d'utilitzar tokens del context amb BoW i sense features.*

Veient que els millors resultats s'han obtingut amb BoW, estenem aquesta idea però aquest cop incloent features per als tokens del context. En aquest cas, utilitzarem els features que apliquen sobre els prefixos i sufixos. Els resultats els podem observar a la Figura 7.3.

Tot i que la millor opció segueix sent seleccionar una paraula, els resultats obtinguts no milloren en relació a l'experiment passat, produint unes puntuacions de 0.63496 pel castellà, i de 0.61286 pel català.

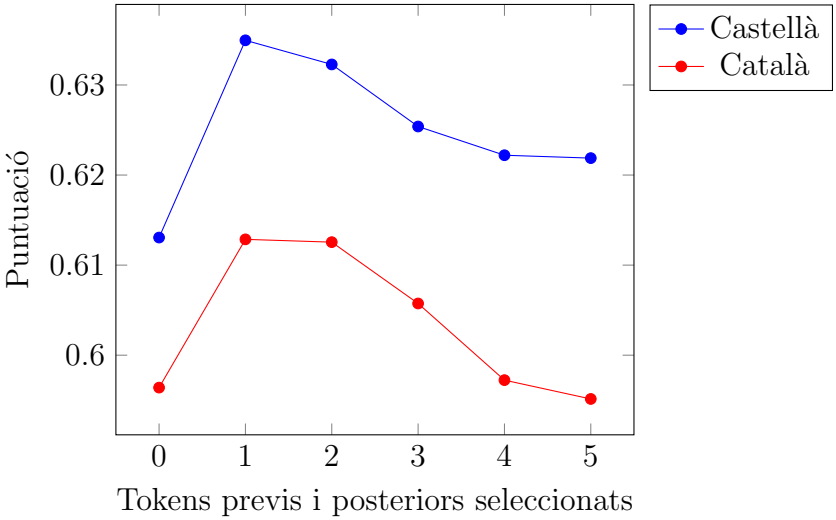


Figura 7.3: *Gràfica dels resultats d'utilitzar tokens del context amb BoW i els features que apliquen sobre els prefixos i sufixos.*

Per finalitzar aquesta experimentació sobre els tokens del context, combinarem les dues opcions: utilitzarem les paraules amb ordre i també BoW. En aquest cas, per les paraules amb ordre aplicarem el mateix conjunt de features que sobre el token actual.

L'experimentació la limitarem a com a màxim 3 paraules, ja que veient els resultats previs podem descartar les opcions de 4 i 5 paraules.

A la Figura 7.4 podem observar que els resultats segueixen la tendència que s'estava produint fins ara, essent el token previ i el posterior la millor elecció. En aquest cas, però, tampoc es milloren els resultats, obtenint una puntuació de 0.63246 pel castellà i 0.61342 pel català.

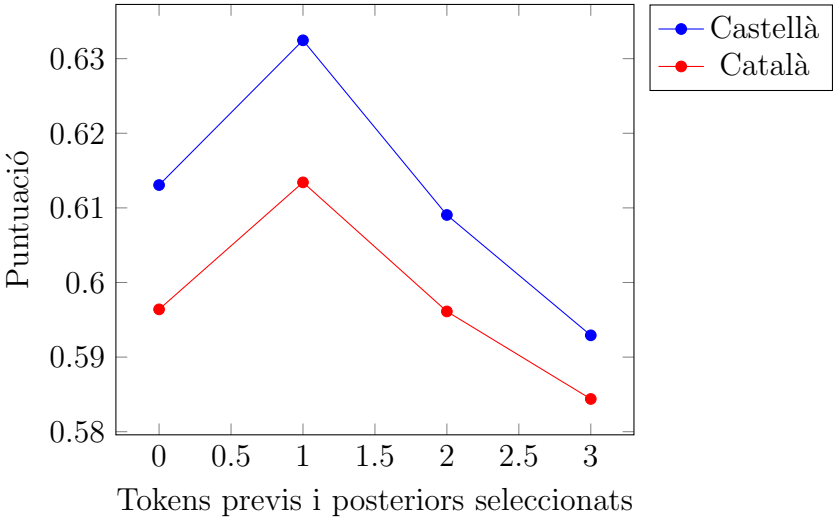


Figura 7.4: *Gràfica dels resultats d'utilitzar tokens del context combinant ordre i BoW.*

Per tant finalitzem l'experimentació sobre la informació del context obtenint els millors resultats aplicant la tècnica de BoW agafant el token previ i el token posterior al que s'està tractant.

### 7.1.2. Esquema d'etiquetatge i model bilingüe

Els objectius d'aquest apartat seran dos: en primer lloc determinar quin és l'esquema d'etiquetatge que millor funciona (BIO/BIOS) i en segon lloc, analitzar quin és el comportament del nostre model quan l'apliquem per documents en català i castellà sense diferenciar-los.

Pel que fa a la primera part, tenim els resultats de l'experimentació a la Taula 7.11.

| | **Castellà** | **Català** |
|---|---|---|
| **BIO** | 0.62807 | 0.60892 |
| **BIOS** | 0.63563 | 0.61444 |

Taula 7.11: *Resultats en funció de l'esquema d'etiquetatge.*

Podem observar que tant pel català com pel castellà els resultats són millors amb l'esquema BIOS. De fet té sentit, ja que si recordem l'anàlisi del capítol *5. Descripció de les dades*, la majoria de les nostres entitats està composta d'una única paraula.

Tot seguit, s'aplica el model sobre tot el conjunt de documents del que es disposa, sense fer la distinció entre tipus de llenguatge. El resultat obtingut d'aquesta aplicació és de 0.64826, una puntuació millor que qualsevol de les obtingudes anteriorment.

Això pot ser conseqüència d'alguns documents en els que s'hi fa una mescla de llenguatges (e.g. el doctor escriu en català però apunta el que el pacient li comenta en castellà) o situacions de l'estil.

A més, hem de recordar que el nostre corpus de dades no és especialment extens, i si ajuntem les dades en català i en castellà, el que estem fent es entrenar i validar el nostre model amb pràcticament el doble de dades de quan ho fem amb la distinció del llenguatge, fet que també contribueix a aquest resultat.

Per tant, en vista a això, podem concloure que el nostre model és bilingüe i tot l'anàlisi que ve posteriorment el farem també per al model sense distinció de llenguatge.

### 7.1.3. Regularització i *folds*

L'algorisme d'optimització que s'ha utilitzat per al nostre model es el LBFGS (*Limited-memory* BFGS), un mètode del tipus quasi-Newton[5] que s'utilitza de manera generalitzada en els CRFs fent regularització.

La regularització en aquests mètodes, és un procés mitjançant el qual s'afegeix informació a l'entrenament per tal d'evitar que el model estigui esbiaixat. Existeixen diferents maneres d'aplicar regularització en sistemes basats en CRFs[18], però en el nostre cas ens basarem en la regularització L1 i L2.

Conseqüentment, l'elecció dels paràmetres de regularització associats a aquest mètodes té un impacte significatiu en el rendiment del model. Per tant, farem un anàlisi per trobar uns paràmetres que en millorin el comportament.

Lògicament és complicat estudiar de manera exhaustiva la totalitat dels paràmetres, per tant el que s'ha fet és una cerca sobre un nombre determinat d'ells guiant-nos pels que vagin millorant els resultats anteriors.

L'experimentació s'ha fet provant 15 combinacions diferents de parelles de paràmetres *c1* i *c2* (associats a les regularitzacions L1 i L2 corresponentment). El rang de possibles valors s'ha acotat a l'escala de 0.1 ja que fent experimentacions prèvies s'ha vist que valors majors eren perjudicials.

[5]Família de mètodes usats per trobar màxims i mínims de funcions. S'utilitzen com a alternativa als mètodes de Newton quan el càlcul de la matriu Jacobina o Hessiana és massa costós computacionalment.

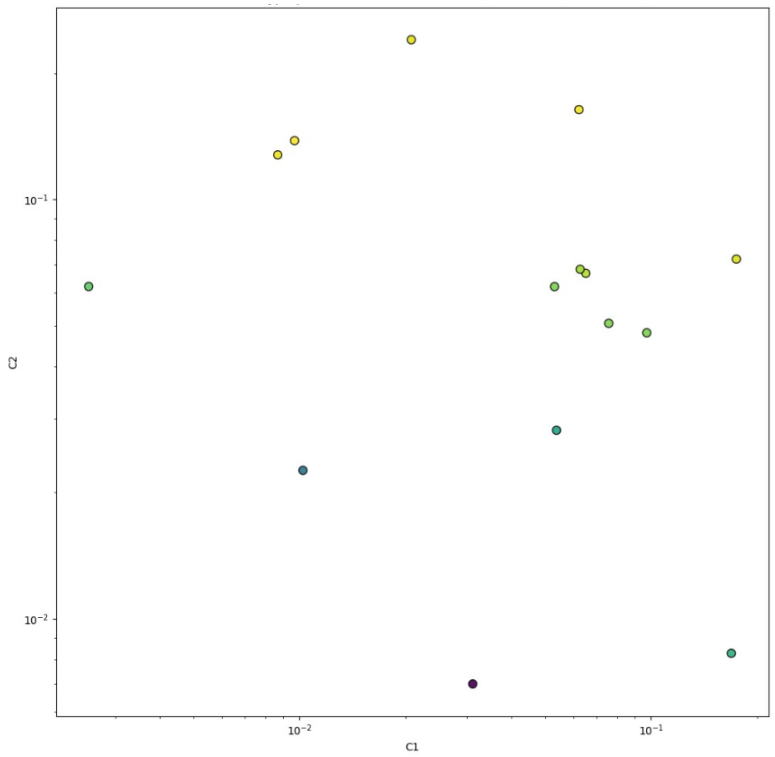


Figura 7.5: *Paràmetres de regularització pel model en castellà.*

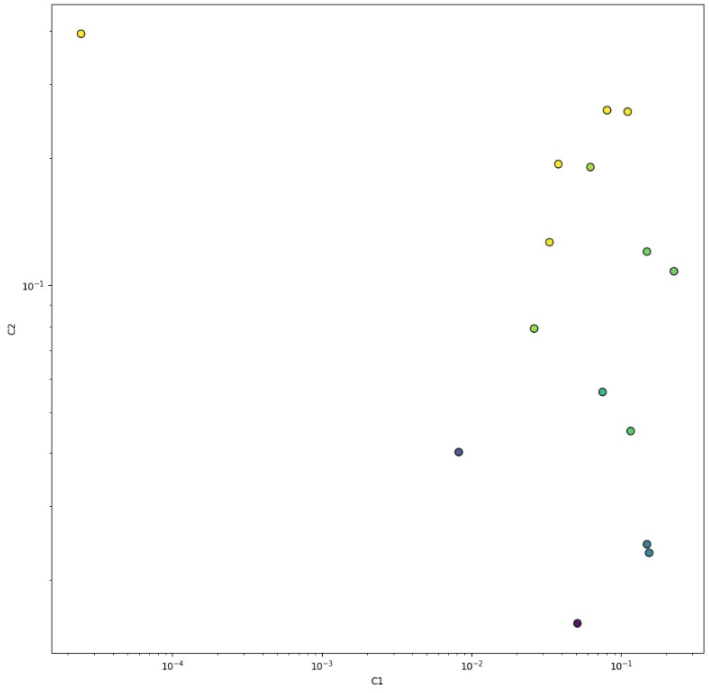


Figura 7.6: *Paràmetres de regularització pel model en català.*

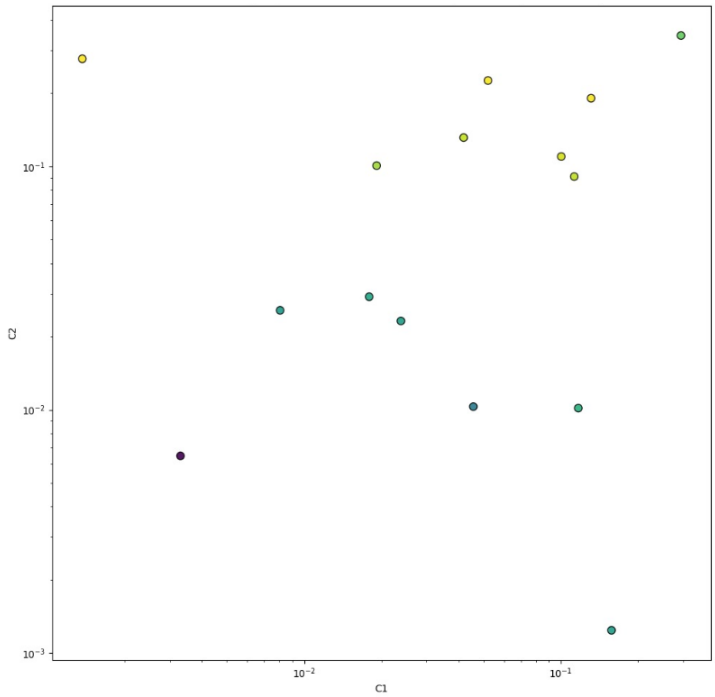


Figura 7.7: *Paràmetres de regularització pel model sense distinció de llenguatge.*

A les Figures 7.5, 7.6 i 7.7 veiem els diferents paràmetres tractats per els models sobre castellà, català i ambdós respectivament. Un mateix punt representa una parella de valors c1,c2 on el color més clar indica unes puntuacions més positives.

Un cop finalitzada l'experimentació, tenim els següents resultats pels corresponents paràmetres, millorant les puntuacions obtingudes fins ara:

- Castellà (c1: 0.00968, c2: 0.13815) → 0.63777
- Català (c1: 0.03805, c2: 0.19374) → 0.61587
- Bilingüe (c1: 0.00137, c2: 0.27702) → 0.65367

Recordem que hem estat aplicant *5-fold Cross Validation* per tant, ara que tenim els millors resultats, s'aporta la puntuació de cada *fold* i la desviació estàndard (Taula 7.12), assegurant que el comportament és consistent en les diferents particions i que per tant els nostres resultats son vàlids.

| | **Fold 1** | **Fold 2** | **Fold 3** | **Fold 4** | **Fold 5** | **Mitjana** | **Desv. est.** |
|---|---|---|---|---|---|---|---|
| **Castellà** | 0.60211 | 0.66491 | 0.63495 | 0.64940 | 0.63744 | **0.63777** | 0.02074 |
| **Català** | 0.62329 | 0.60328 | 0.60706 | 0.62300 | 0.62269 | **0.61587** | 0.00881 |
| **Bilingüe** | 0.66515 | 0.65894 | 0.63360 | 0.65253 | 0.65803 | **0.65367** | 0.01079 |

Taula 7.12: *Resultats per cada fold i desviació estàndard.*

| | **Precisió** |
|---|---|
| **Castellà** | 0.93299 |
| **Català** | 0.94180 |
| **Bilingüe** | 0.94072 |

Taula 7.13: *Puntuació de precisió contemplant la totalitat dels tokens.*

I finalitzem mostrant a la Taula 7.13 la puntuació basada en la precisió que hem tingut per predir cada token, incloent aquells que s'etiqueten com a no pertanyents a cap categoria (O). Òbviament les puntuacions són molt elevades, ja que la majoria de tokens estan classificats com a no pertanyents a cap de les categories.

No és d'estranyar veure que obtenim millors puntuacions en models que abans donaven pitjors, ja que aquí estem avaluant token per token i per tant es pot donar el cas d'una entitat composta de més d'un token, on algun token s'ha predit bé i d'altres malament. Aquest cas es consideraria error per la f1-score per entitat, i en canvi si que hi hauria encerts per la precisió sobre el total.

### 7.1.4. Puntuacions per categoria

Amb l'objectiu d'analitzar els resultats en major detall, estudiarem les puntuacions de precisió, recall i f1-score desfragmentades segons la categoria en la que es classifica.

A més es mostren també la *macro*[6] i la *micro average*[7] totals de les puntuacions. Els resultats s'exposen a les Taules 7.14, 7.15 i 7.16.

| | **Precisió** | **Recall** | **F1-Score** |
|---|---|---|---|
| **Fàrmac** | 0.930 | 0.665 | 0.776 |
| **Part del Cos** | 0.634 | 0.415 | 0.502 |
| **Signe / Símptoma** | 0.713 | 0.579 | 0.639 |
| **Diagnòstic** | 0.715 | 0.507 | 0.593 |
| | | | |
| **Micro avg.** | 0.752 | 0.554 | 0.638 |
| **Macro avg.** | 0.748 | 0.542 | 0.628 |

Taula 7.14: *Puntuacions fragmentades per classes del model en castellà.*

[6]Càlcul de la mètrica independentment per cada classe i posteriorment fa la mitjana.

[7]Com l'anterior, però tenint en compte les contribucions de cada classe a l'hora de calcular la mitjana.

| | Precisió | Recall | F1-Score |
|---|---|---|---|
| **Fàrmac** | 0.923 | 0.677 | 0.781 |
| **Part del Cos** | 0.629 | 0.386 | 0.479 |
| **Signe / Símptoma** | 0.653 | 0.499 | 0.566 |
| **Diagnòstic** | 0.724 | 0.524 | 0.608 |
| | | | |
| **Micro avg.** | 0.735 | 0.528 | 0.616 |
| **Macro avg.** | 0.732 | 0.523 | 0.609 |

Taula 7.15: *Puntuacions fragmentades per classes del model en català.*

| | Precisió | Recall | F1-Score |
|---|---|---|---|
| **Fàrmac** | 0.939 | 0.724 | 0.817 |
| **Part del Cos** | 0.652 | 0.402 | 0.497 |
| **Signe / Símptoma** | 0.705 | 0.560 | 0.624 |
| **Diagnòstic** | 0.738 | 0.559 | 0.637 |
| | | | |
| **Micro avg.** | 0.764 | 0.570 | 0.653 |
| **Macro avg.** | 0.758 | 0.561 | 0.644 |

Taula 7.16: *Puntuacions fragmentades per classes del model bilingüe.*

En primer lloc, destacar el bon comportament que té el model en els tres casos en la detecció dels Fàrmacs. Especialment cal remarcar les precisions tant elevades en les que ens movem, fet que indica que dels elements marcats com a Fàrmacs, la immensa majoria ho són realment i per tant tenim molt pocs falsos positius.

Aquest bon comportament en relació als Fàrmacs té dos elements com a causants principals:

- Primerament i el més important, hem vist que els fàrmacs són, amb bastant diferència, la categoria amb major nombre d'entitats d'un sol token (95%), i això pot simplificar-ne la detecció (amb la matriu de confusió ho corroborem).
- D'altra banda, acostumen a ser aquelles entitats que s'escriuen d'una manera peculiar, amb uns prefixos i sufixos molt determinats, fet que facilita la seva identificació.

Seguint amb l'anàlisi, veiem que la categoria amb la que obtenim unes pitjors puntuacions és amb les Parts del Cos. En especial destaquem la puntuació de recall, fet que ens fa veure que del total d'elements que es consideren parts del cos, n'estem seleccionant una proporció força baixa.

Aquest baix rendiment detectant aquelles entitats que són parts del cos pot ser causat perquè no estem definint bé les seves característiques o no estem aprofitant bé el context. És un defecte que es podria tractar de suplir elaborant un llistat extraient coneixement d'experts, on es poguessin emmagatzemar parts del cos per així poder detectar-les amb major facilitat.

### 7.1.5. Matriu de confusió

En aquesta secció, elaborarem les matrius de confusió per analitzar de quina manera específica s'estan s'està produint l'etiquetatge, i poder veure més a fons quins són els punts forts i punts febles del nostre model.

La matriu es mostra a les Figures 7.8, 7.9 i 7.10 pel castellà, català i model bilingüe respectivament. L'eix d'ordenades indica les etiquetes correctes (es a dir, com apareixen anotades al nostre corpus) i l'eix d'abscisses mostra les etiquetes que ha associat el model.

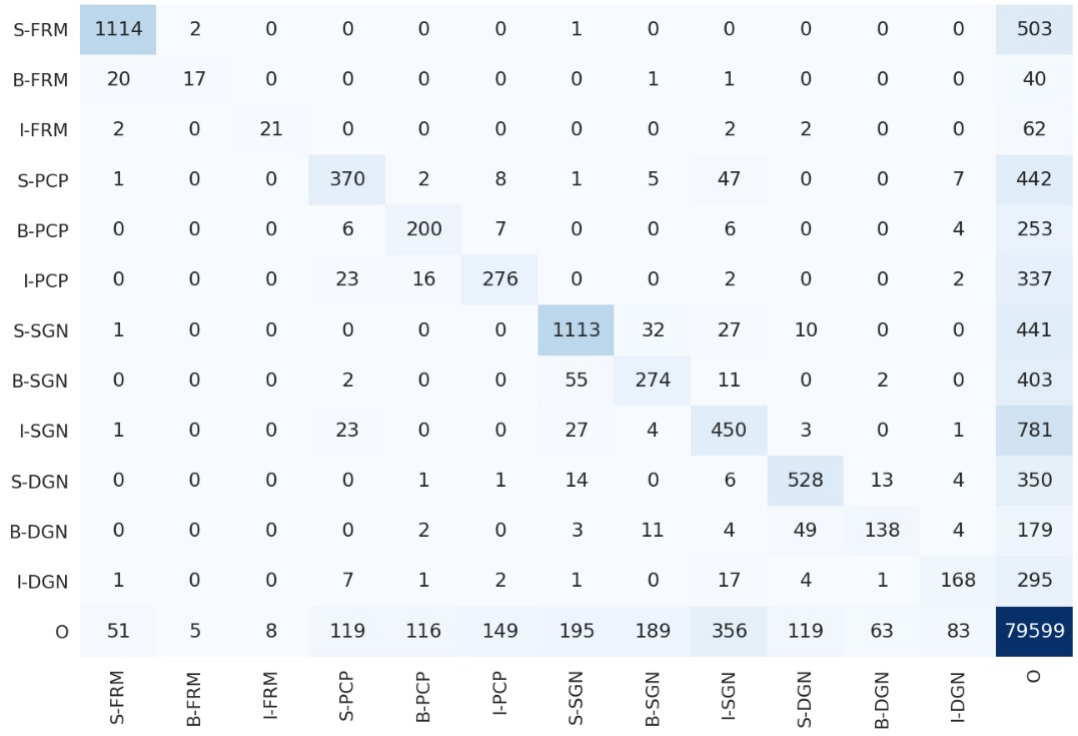


Figura 7.8: *Matriu de confusió pel model en castellà.*

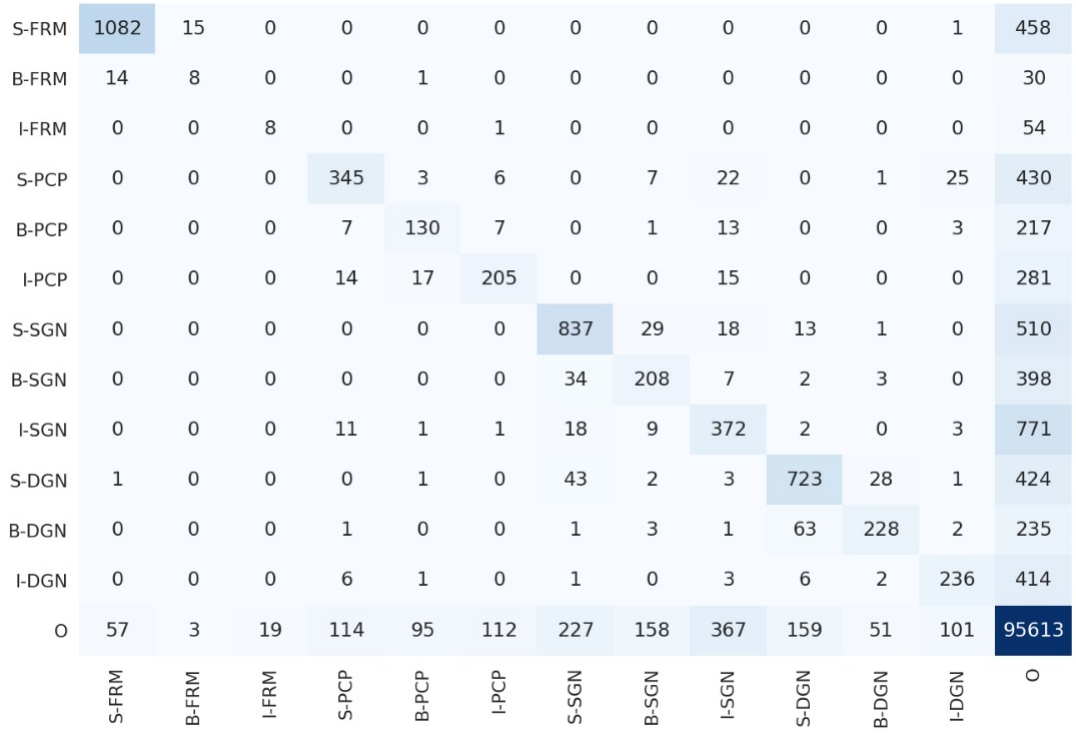


Figura 7.9: *Matriu de confusió pel model en català.*

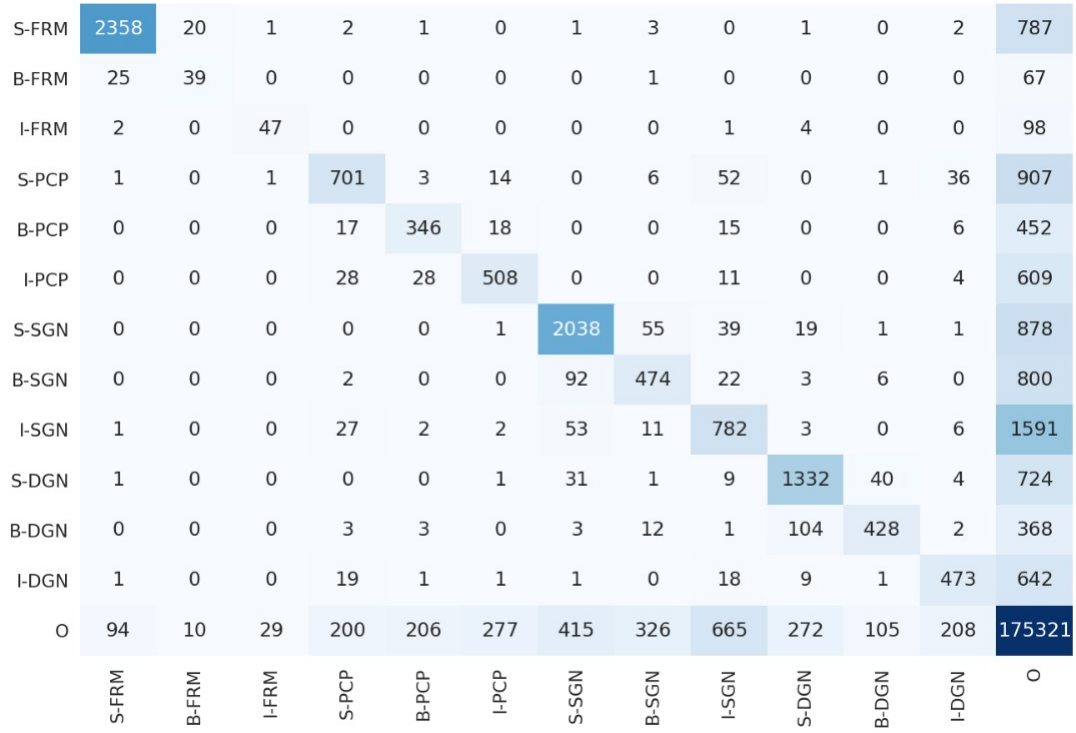


Figura 7.10: *Matriu de confusió pel model sense distinció de llenguatge.*

De la matriu de confusió, la primera característica que es reprodueix als tres models i que podem identificar, és una confusió general entre els etiquetatges d'una mateixa categoria quan comencen per "B-" i s'etiqueten amb "S-" i al revés.

De fet, és una confusió que en major o menor mesura, s'espera en models d'etiquetatge de seqüències utilitzant l'esquema BIOS. El que succeeix és que s'estan etiquetant entitats de més d'una paraula com si només en fos una (llavors el la primera paraula és del tipus "B-" i s'etiqueta amb "S-" i a la inversa.

També veiem en els tres models les Parts del Cos, i en especialment les entitats S-PCP que són una part del cos composta per una única paraula, com s'acostumen a confondre amb d'altres (especialment amb I-SGN i I-DGN).

Aquesta confusió rau en que situacions com "*càncer pulmonar*"al corpus són etiquetades com "*S-Diagnòstic S-Part del Cos*" i en el nostre cas, ho estaríem considerant tot com a diagnòstic "*B-Diagnòstic I-Diagnòstic*".

Generalment, aquesta segona part de l'entitat acostuma a determinar la localització a la que fa referència la primera part de l'entitat (sigui aquesta Diagnòstic o Símptoma) i és per això que tenim aquest tipus de confusió.

### 7.1.6. Exemples d'errors de predicció

Per finalitzar l'experimentació, ens centrarem en trobar exemples reals d'errors de predicció que mostrin o justifiquin els comportaments que hem anat caracteritzant.

En primer lloc, hem detectat que el nostre model detecta Fàrmacs com a falsos positius (Veure puntuacions de recall *7.1.4 Puntuacions per categoria*) quan aquestes paraules tenen un patró similar (generalment pel tipus de prefix o de sufix) al dels fàrmacs reals.

Un exemple d'això seria el següent, on *portin* que no ha de pertànyer a cap categoria, l'hem classificat com a Fàrmac per la seva semblança a altres (e.g. *neurotin, lexatin, loratin*).

"*...per lo que s'avisa a la família perquè el* ***portin***"

Seguim amb un exemple del que s'ha comentat a la secció anterior en relació a la confusió dels etiquetatges amb S-Part del Cos. En aquest cas tenim un exemple en el que *infecció bronquial* s'etiqueta al corpus com *S-Diagnòstic S-Part del cos* i el nostre model en canvi, etiqueta com *B-Diagnòstic I-Diagnòstic*.

"*...Durant el seu ingrés hospitalari va fer* ***infecció bronquial*** *que es va resoldre amb...*"

Per finalitzar, comentar també que ens hem trobat amb diferents entitats que a les dades no es trobaven anotades però que el nostre model sí que les anota i considerem que fer-ho és el comportament correcte.

En el següent exemple, *doloroso a la palpación* no es contempla a cap de les categories, però el sistema ho etiqueta com a Signe / Símptoma.

"*...depresible,* ***doloroso a la palpación*** *en flanco derecho*"

## 7.2. Mètode 2: Bi-LSTM-CRF

En aquesta secció ens centrarem en l'anàlisi del segon mètode, analitzant la seva configuració i comportament i avaluant els seus resultats.

Seguint l'esquema de l'experimentació amb CRFs, hem utilitzat un 80% de les dades pel procés d'entrenament, i un 20% pel de test. Per les puntuacions també experimentarem basant-nos en la f1-score, i més endavant mostrarem precisió i *recall*.

### 7.2.1. *Embeddings* i llenguatge

Comencem l'experimentació analitzant el comportament del nostre model en funció del tipus de word embeddings utilitzats: Fasttext o Word2Vec. Aquest comportament l'analitzarem tant fent la distinció de llenguatges en català i castellà com sense aquesta distinció.

Abans, però, de començar a exposar resultats, es fa un llistat d'alguns dels paràmetres que s'han escollit per la configuració inicial de la xarxa neuronal:

- *Trainable embeddings*[8]
- Dropout: 0.1
- Recurrent Dropout: 0.1
- Funció d'activació: relu
- Optimitzador: rmsprop
- Batch size: 64

[8]Fem que els pesos puguin set actualitzats durant l'entrenament, adaptant els embeddings al la forma del nostre corpus.

- Epochs: 10
- Split de validació: 0.2 (dins del 80% de dades d'entrenament)

En base a això, veiem els resultats de l'experimentació del tipus d'embeddings i llenguatge a la Taula 7.17.

| | **Fasttext** | **Word2Vec** |
|---|---|---|
| **Castellà** | 0.67589 | 0.65199 |
| **Català** | 0.64625 | 0.62662 |
| **Bilingüe** | 0.68695 | 0.65925 |

Taula 7.17: *Puntuacions en funció del tipus de word embedding i llenguatge.*

En primer lloc podem veure que es segueixen les tendències en relació al tipus de llenguatge usat, essent el model bilingüe el que aporta millors resultats (recordem que segurament es deu a mescles d'idiomes en el corpus i major quantitat de dades disponibles per entrenar).

I en segon lloc, podem confirmar la nostra hipòtesi inicial en relació al tipus d'embedding, i és que clarament veiem un millor comportament amb el tipus Fasttext que amb el Word2Vec.

Això és gràcies a que els Fasttext tracten n-grames i no paraules com a unitat atòmica, i això aporta un millor comportament tractant amb paraules escrites amb equivocacions i faltes ortogràfiques.

Finalment, és interessar veure que estem obtenint unes puntuacions clarament superiors al mètode basat en CRFs, fet que ja analitzarem en profunditat més endavant.

Per tant, en base a aquests resultats, utilitzarem per l'experimentació que segueix els embeddings de tipus Fasttext i el model bilingüe sense distinció de llenguatge.

### 7.2.2. Hiperparàmetres

Els hiperparàmetres són aquells paràmetres que es defineixen abans del procés d'entrenament i que generalment, és difícil de terminar quins seran els millors sense una experimentació prèvia.

Per tant, en aquesta secció ens centrarem en explorar-ne alguns d'ells seleccionant aquells que millorin el rendiment del nostre sistema.

Per començar ens centrarem en el *batch size*. Aquest paràmetre ens indica quin és el nombre d'exemples de dades (en el nostre cas, nombre de documents) que conformen una passada per la xarxa neuronal. Per a cada *epoch*[9], seran necessàries tantes passades com *batch size* calgui sumar fins a completar la quantitat de dades d'entrenament.

Aquest, acostuma a ser un factor important en el rendiment d'una xarxa neuronal, i per tant serà per on comencem.

A banda dels resultats que s'obtenen per als diferents valors (Taula 7.18), serà interessant veure les gràfiques de l'entrenament, observant els valors de les mètriques *accuracy* i *loss* obtingudes al llarg de les diferents epoch en els conjunts d'entrenament i validació. Aquestes gràfiques les podem veure representades a les figures següents.

| | **32** | **64** | **128** | **256** |
|---|---|---|---|---|
| **f1-score** | 0.68256 | 0.68695 | 0.68296 | 0.67772 |

Taula 7.18: *Puntuacions per els diferents valors de batch size.*

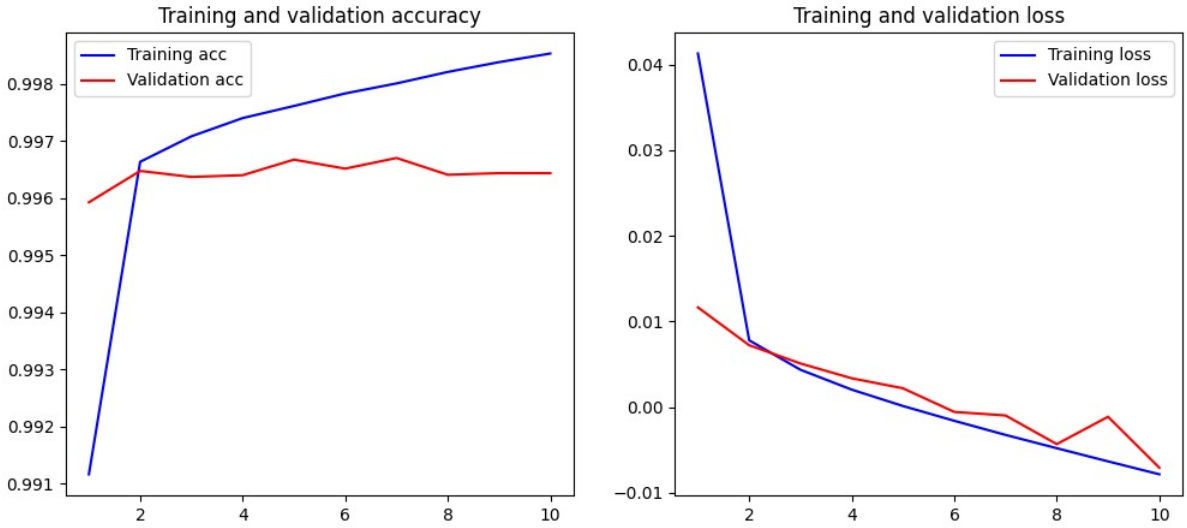


Figura 7.11: *Gràfica del procés d'entrenament amb un batch size de 32.*

[9]Iteració en la que es passen totes les dades d'entrenament per la xarxa neuronal un cop.

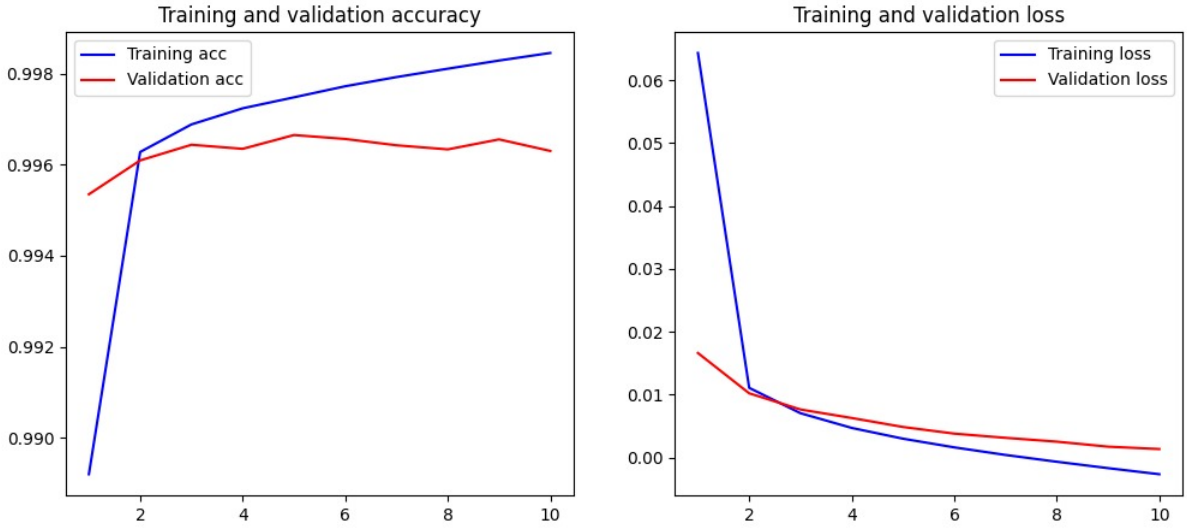


Figura 7.12: *Gràfica del procés d'entrenament amb un batch size de 64.*

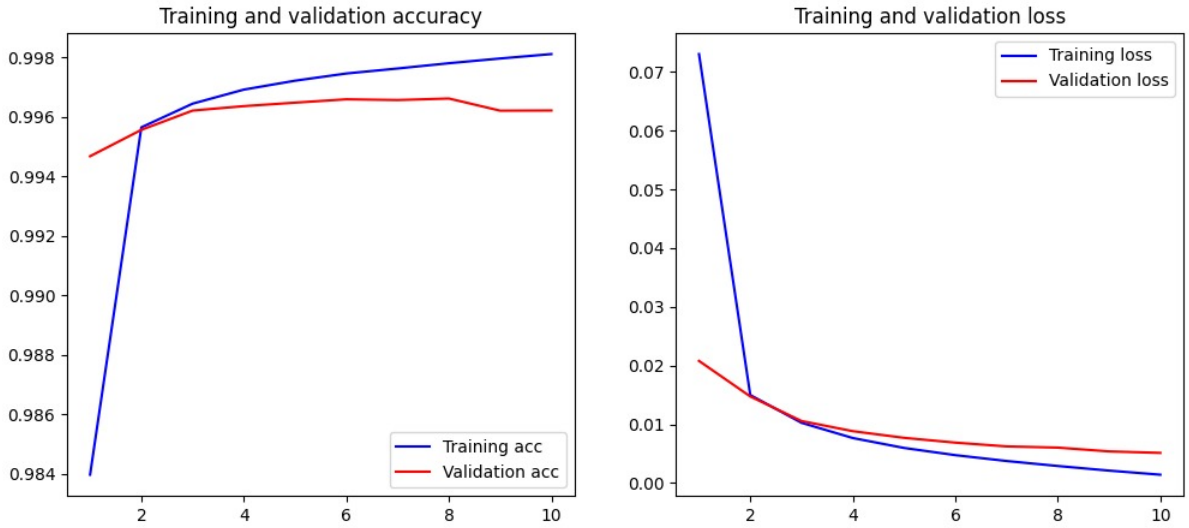


Figura 7.13: *Gràfica del procés d'entrenament amb un batch size de 128.*

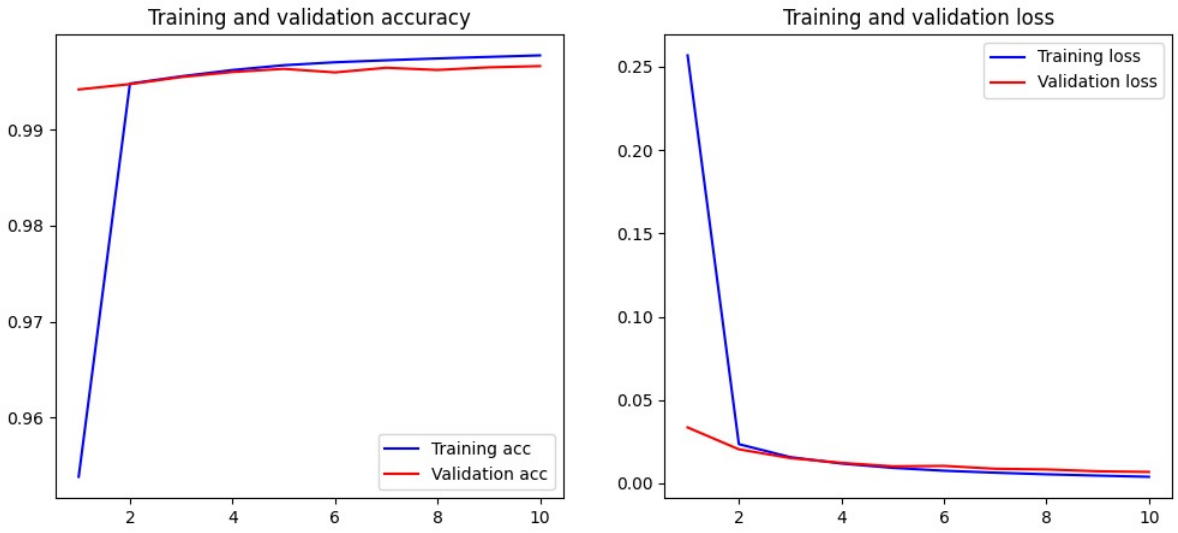


Figura 7.14: *Gràfica del procés d'entrenament amb un batch size de 256.*

És interessant veure el progrés durant l'entrenament, on podem observar que a mesura que augmenta el batch size, el model convergeix abans.

En quant a resultats, veiem que el millor valor s'obté amb 64, i serà per tant aquest valor el que mantindrem durant la resta d'experimentació. A més, no interessa tenir un batch size molt elevat, ja que el nostre model estaria perdent nivell de generalització.

Seguim l'experimentació amb els optimitzadors, que són algoritmes que modifiquen els atributs de la nostra xarxa neuronal durant l'entrenament. El conjunt d'algoritmes que provarem s'ha seleccionat pensant en que poden ser bons candidats per a problemes de NER utilitzant xarxes neuronals recurrents, i és el següent:

- RMSprop
- SGD (*Stochastic Gradient Descent*)
- Adam
- AdaGrad

Els resultats els podem observar a la Taula 7.19, i a més també disposem dels gràfics per a presentar el procés d'entrenament.

| | **RMsprop** | **SGD** | **Adam** | **AdaGrad** |
|---|---|---|---|---|
| **f1-score** | 0.68695 | 0.0 | 0.66880 | 0.63397 |

Taula 7.19: *Puntuacions per els diferents optimitzadors.*

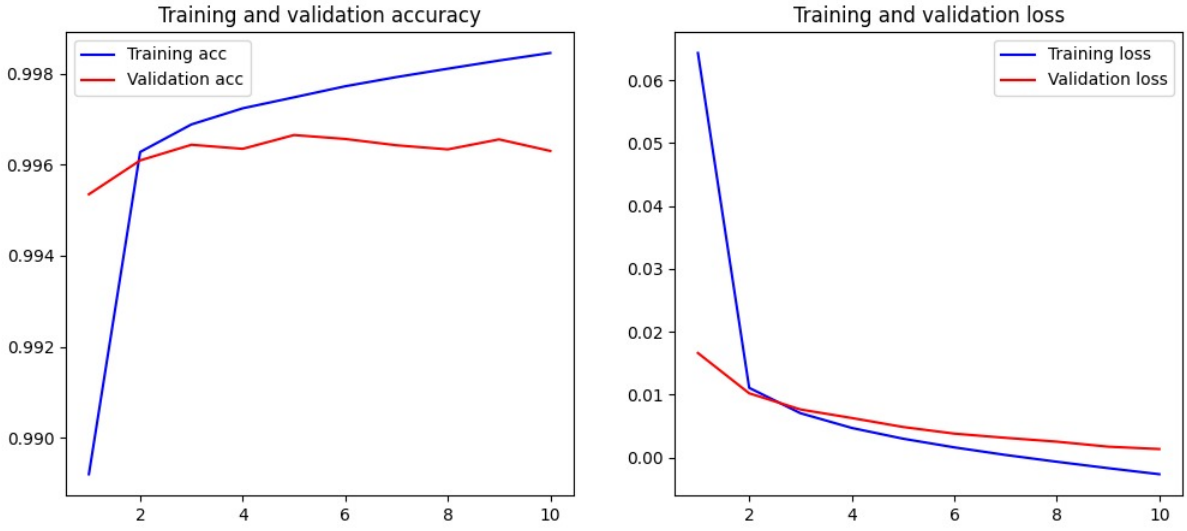


Figura 7.15: *Gràfica del procés d'entrenament amb l'optimitzador RMsprop.*

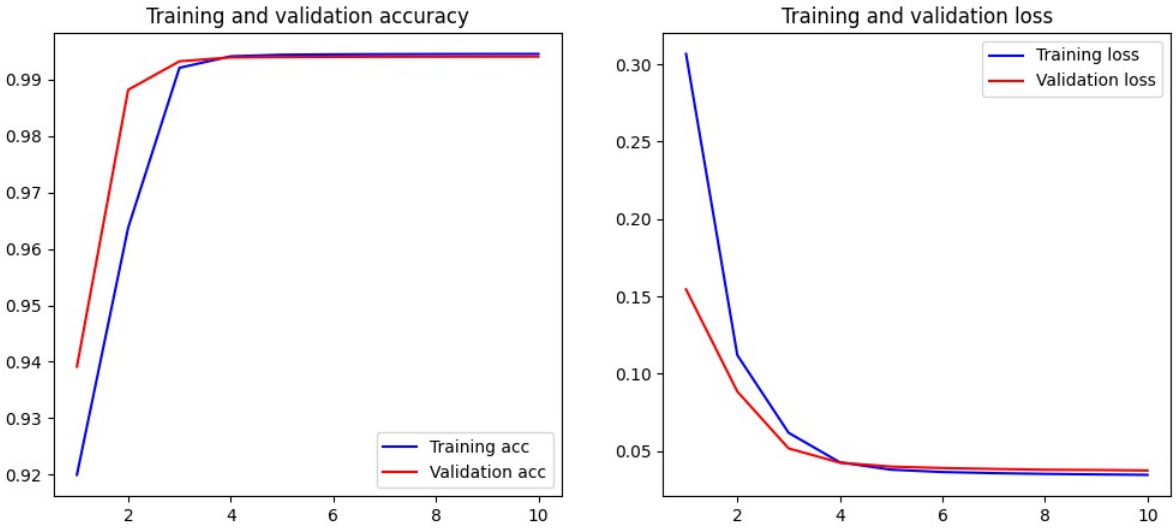


Figura 7.16: *Gràfica del procés d'entrenament amb l'optimitzador SGD.*

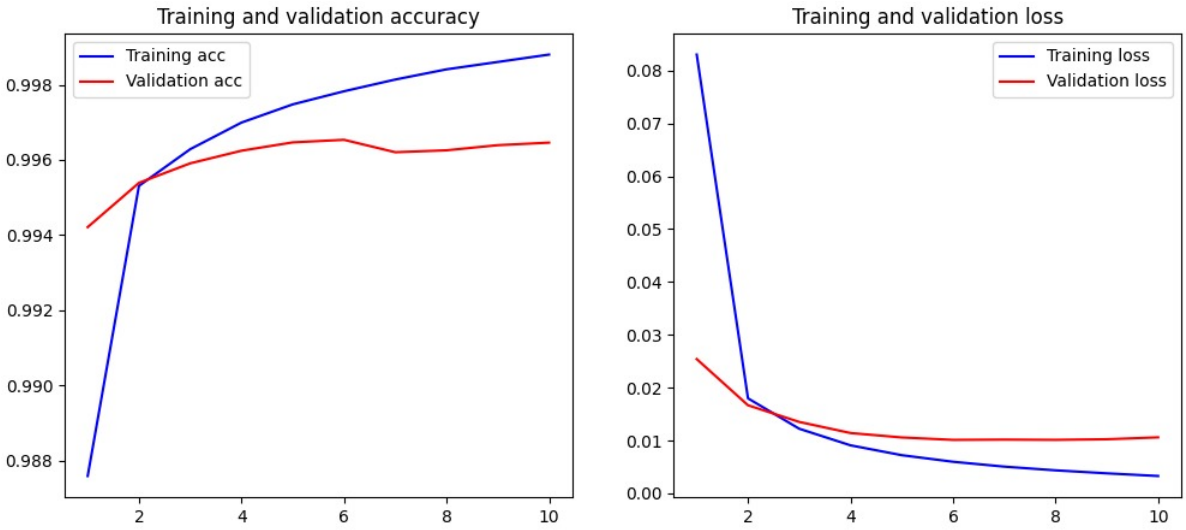


Figura 7.17: *Gràfica del procés d'entrenament amb l'optimitzador Adam.*

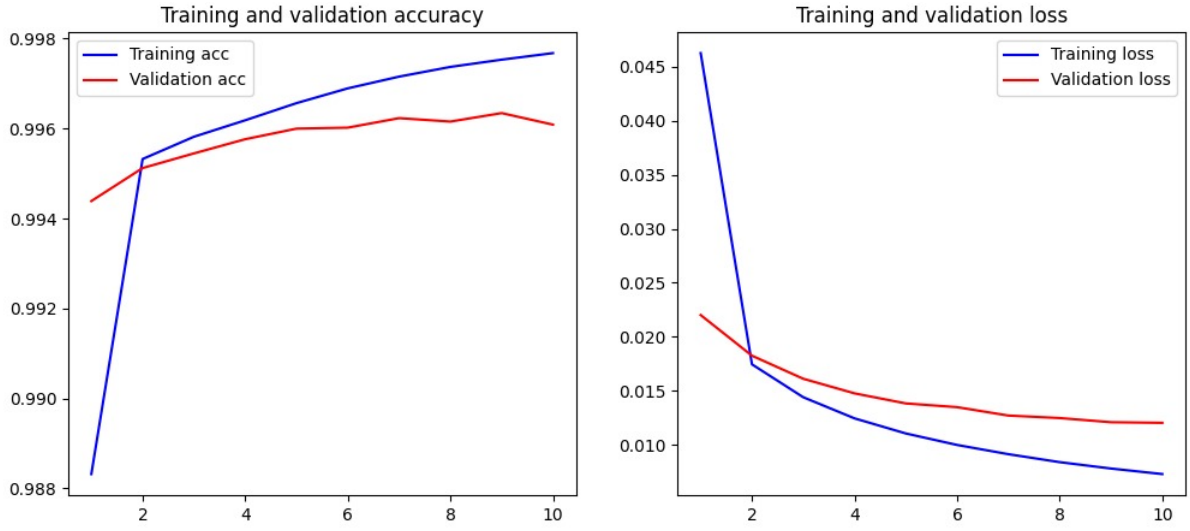


Figura 7.18: *Gràfica del procés d'entrenament amb l'optimitzador AdaGrad.*

Dels resultats en poder obtenir diferents conclusions. En primer lloc, descartem directament l'optimitzador SGD. Aquest resultat de 0 és causat perquè s'encalla en un màxim local, etiquetant totes les entitats amb l'etiqueta "O", fet que fa que s'obtingui al voltant d'un 0.90 de puntuació en precisió total, però un 0 en entitats anotades correctament. De fet, a la Figura 7.16 és pot veure gràficament com el model s'encalla.

Seguint amb l'anàlisi, veiem que el comportament del AdaGrad tampoc millora, encara que aquest indicador acostuma a tenir un bon comportament en tasques de NPL.

L'optimitzador Adam, que combina característiques del RMSprop i del AdaGrad obté ja uns resultats més interessants, però els del RMSprop segueixen sent millors, essent aquest un optimitzador indicat per a treballs amb xarxes neuronals recurrents.

Vists els resultats, i per tal de no escollir sempre la millor opció i poder arribar així a un màxim local, ens quedarem per a la següent experimentació els optimitzadors RMSprop i Adam.

Per finalitzar l'experimentació dels hiperparàmetres, estudiarem quines funcions d'activació funcionen millor. Aquestes, són les funcions que defineixen la sortida d'un node en funció de les seves entrades.

La selecció de funcions d'activació s'ha fet en base a les que acostumen a aportar millors resultats. El conjunt de funcions serà el següent:

- SoftMax
- ReLU
- Sigmoid

Els resultats els podem observar a la Taula 7.20. on fàcilment podem veure que la funció ReLU és la que aporta millors resultats tant per RMSprop com per Adam. Destaquem també que amb SoftMax també arribem a un màxim local on s'etiqueta tot amb 'O' (excepte per un *Fold* amb RMSprop).

| | **SoftMax** | **ReLU** | **Sigmoid** |
|---|---|---|---|
| **RMSprop** | 0.09772 | 0.68695 | 0.67335 |
| **Adam** | 0.0 | 0.66880 | 0.56251 |

Taula 7.20: *Puntuacions per les diferents funcions d'activació.*

### 7.2.3. *Folds*

Recordem que estem aplicant *5-Fold Cross Validation*, per tant, ara que tenim els millors resultats, aportarem la puntuació de cada *fold* i la desviació estàndard total. Aquestes dades es mostren a la Taula 7.21, assegurant que el comportament és força consistent entre les diferents particions i que els resultats són vàlids.

| **Fold 1** | **Fold 2** | **Fold 3** | **Fold 4** | **Fold 5** | **Mitjana** | **Desv. est.** |
|---|---|---|---|---|---|---|
| 0.68691 | 0.67304 | 0.68910 | 0.68755 | 0.69817 | **0.68695** | 0.00805 |

Taula 7.21: *Anàlisi de resultats per fold.*

A més, comentar que la puntuació que s'obté de precisió comptant totes les etiquetes, és de 0.94396.

### 7.2.4. Puntuacions per categoria

Tot seguit, a la Taula 7.22 mostrem els resultats desglossats per categoria. Per a cada una, mostrarem a més de la f1-score, la precisió i el recall i també la *micro* i *macro average*.

| | **Precisió** | **Recall** | **F1-Score** |
|---|---|---|---|
| **Fàrmac** | 0.908 | 0.811 | 0.862 |
| **Part del Cos** | 0.614 | 0.667 | 0.639 |
| **Signe / Símptoma** | 0.622 | 0.651 | 0.631 |
| **Diagnòstic** | 0.647 | 0.569 | 0.608 |
| **Micro avg.** | 0.700 | 0.678 | 0.687 |
| **Macro avg.** | 0.698 | 0.674 | 0.685 |

Taula 7.22: *Puntuacions fragmentades per classes.*

Podem observar algunes tendències semblants a les que ens aportava el primer mètode, on les entitats classificades com a fàrmacs són les que millors resultats donen, destacant-ne l'elevada precisió.

En canvi, en aquest cas, l'entitat amb la que tenim majors problemes per categoritzar-la són els Diagnòstics, on tenim un recall força baix, és a dir, ens estem deixant molts elements que són Diagnòstics per anotar.

### 7.2.5. Matriu de confusió

De manera similar a com vam fer amb l'experimentació del primer mètode, el que ara farem és mostrar la matriu de confusió (Figura 7.19) per tal d'analitzar més a fons com esta fent l'etiquetatge el nostre model.

Recordem que l'eix d'ordenades indica aquelles etiquetes correctes inicialment, i l'eix d'abscisses mostra quines són les etiquetes que ha associat el nostre sistema.

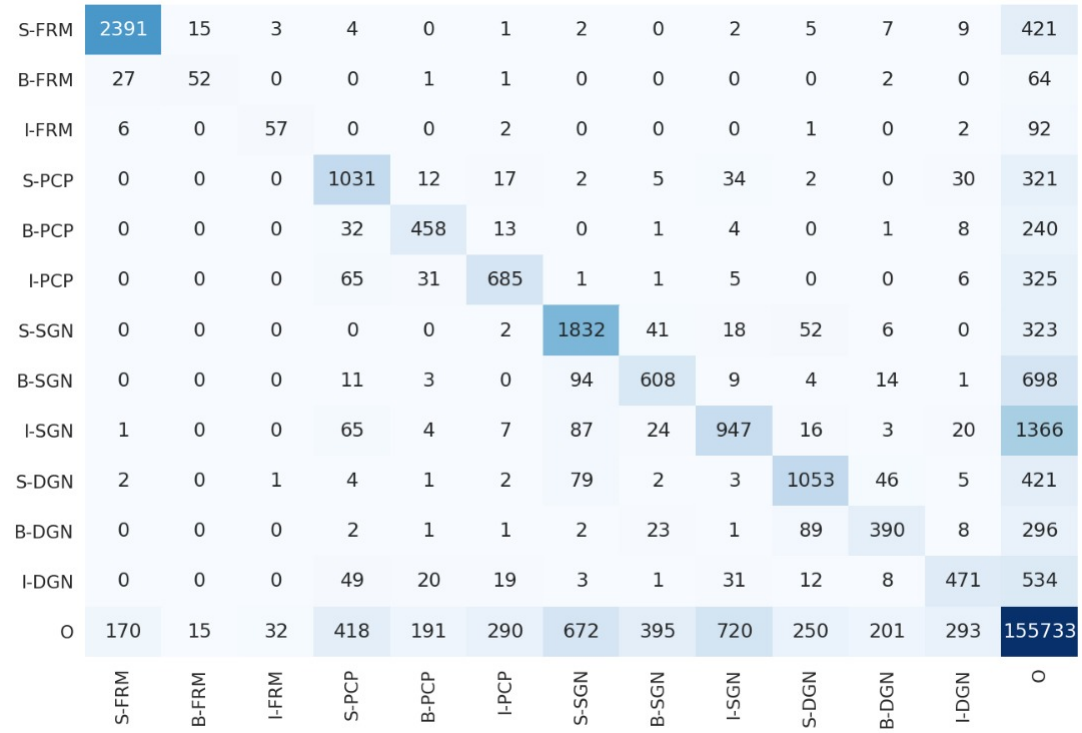


Figura 7.19: *Matriu de confusió.*

D'entre les característiques que hi podem veure, en primer lloc destacar que seguim mantenint la confusió entre termes amb etiquetes que comencen per "S-" i "B-"en els dos sentits. Recordem que això és causat per aquelles entitats que es detecten com a un sol token ("S-") i realment són més (primer token, "B-") i a l'inrevés.

Tot i que ja vam comentar que utilitzant l'esquema d'etiquetatge BIOS aquest és un problema relativament esperable, el que hem fet és analitzar els nostres resultats amb la intenció de detectar seqüencies d'etiquetatge de l'estil "*O - B-Fàrmac - O*"per a aplicar-hi un postprocés i deixar-les correctament etiquetades (*"O - S-Fàrmac - O"*).

No obstant no ens hem trobat casos d'aquest estil (en part, gràcies a la darrera capa de CRF de la xarxa neuronal) i per tant aquest postprocés no comporta cap afectació als resultats.

També veiem, si ens fixem en la columna de S-PCP, que es segueixen etiquetant moltes entitats amb aquesta etiqueta quan realment no ho són, fet que ja vam veure i exposar a l'experimentació sobre CRFs.

És destacable també la confusió entre les entitats que són Signes/Símptomes i que s'etiqueten com a Diagnòstic i al revés, confusió que es produeix al mateix nivell a les corresponents etiquetes del tipus "S-", "B-" i "I-".

Aquesta confusió és en part inherent a l'ambigüitat del llenguatge natural, on en funció del context del document o de com s'interpreti, algunes paraules poden ser considerades símptomes o diagnòstic (e.g. *cefalea*).

### 7.2.6. Exemples d'errors

Darrerament, s'exposaran alguns errors de predicció significatius per tal de definir amb exemples reals algunes de les característiques del nostre model de predicció.

En primer lloc, veure que tenim estem deixant d'etiquetar algunes entitats que tenen patrons molt clars (e.g prefixos i sufixos) i que conseqüentment amb el primer mètode no acostumàvem a fallar (encara que sovint produïa falsos positius).

Un exemple d'això seria el que veiem a continuació, on *neumonia nosocomial greu* no s'ha etiquetat, i té un patró de Diagnòstic per la seva terminació o per anar acompanyat de la paraula *greu*.

*"...Es* ***descarta neumonia nosocomial*** *greu."*

De manera similar al que veiem amb el primer mètode, seguim trobant algunes entitats que no estan etiquetades i que segons el nostre model (i el nostre criteri) si que ho haurien d'estar.

Per exemple, en aquest cas veiem que *Fibromialgia* no s'ha etiquetat dins de cap de les categories, però el nostre model ho etiqueta com a Signe/Símptoma.

*"... Diverticulosis izda por colonoscopia,* ***Fibromialgia****, angioma hepatico ..."*

És interessant veure que aquest terme també ens val com a exemple d'ambigüitat entre Signe/Símptoma i Diagnòstic.

# 8. Conclusions

El reconeixement i la classificació d'entitats mèdiques en els informes clínics és un problema on la major dificultat rau en com adaptar els models al característic tipus de llenguatge que s'utilitza en els documents mèdics.

En aquest sentit, els resultats del mètode basat en CRFs queden lluny dels que s'esperen en relació a l'estat de l'art. En canvi, amb l'alternativa basa en xarxes neuronals Bi-LSTM-CRF obtenim uns resultats significativament millors, i això ens aporta una sèrie de conclusions molt interessants de cara a futures línies d'investigació:

- La primera alternativa aporta uns nivells de precisió lleugerament més elevats, però en canvi, les puntuacions de recall a totes les classes són molt més baixes que les que s'obtenen amb el segon mètode. En base a això, veiem que l'ús dels *word embeddings* basats en n-grames permeten detectar molt millor aquelles entitats escrites amb canvis ortogràfics en relació a la paraula correcta.
- En textos on el canvi de context és molt freqüent i és de molta utilitat, la utilització de LSTMs bidireccionals és interessant en aporta un millor comportament gràcies a l'aprofitament del context previ i posterior.
- L'ús dels *word embeddings*, ens ha permès dotar de molt més coneixement al segon mètode, i això ha tingut clarament un impacte positiu, per exemple, en la detecció de parts del cos, on veiem la major diferencia de resultats entre les dues alternatives. A més, amb els embeddings, aportem més coneixement d'expert al model i augmentem així la seva capacitat de generalització.

Tot i això, tenim una carència comú en els dos models, aquesta són les ambigüitats que trobem entre termes, especialment entre Diagnòstics i Símptomes. En aquest context, línies d'investigació futures basades en la desambiguació d'entitats mèdiques poden contribuir en gran mesura a la qualitat dels models.

A més d'això, hem de considerar que hem estat treballant sobre un corpus de dades relativament reduït (especialment per les xarxes neuronals LSTM), on hem pogut veure que al final els millors resultats s'obtenien quan es podien entrenar els models amb major quantitat de dades sense separar per llenguatge. Per tant, de cara a futur treball sobre aquest camp específic, un augment de la qualitat i el corpus de les dades serien clau per a millorar els resultats i extreure'n millors conclusions.

Un altre aspecte a comentar és la reduïda quantitat de recursos que existeix per aquest problema (i en general pel NLP) per al català en relació del castellà. En vers a això, un interessant futur projecte seria l'elaboració d'uns *word embeddings* basats en la ontologia mèdica per la llengua catalana, fet que juntament amb un major nombre de dades, ens permetria probablement poder fer distinció entre els llenguatges, podent tractar més a fons les peculiaritats de cadascun.

En conclusió, la investigació en aquest camp encara té força recorregut a fer ja que són molts els aspectes que s'han de tractar a fons i que encara és difícil de fer-ho amb les tècniques i recursos disponibles avui en dia. De fet, això és vàlid en general pel propi camp del NPL, que tot i que avança a un gran ritme, encara són moltes les problemàtiques a enfrontar i solucionar lluitant contra la complexitat inherent al llenguatge natural.

# 9. Referències


[1] Ixa taldea - Grup investigació UPV. PROSAMED 2019 [en línia]. [Consulta 25 març 2020]. Disponible a http://ixa2.si.ehu.eus/prosamed/es.

[2] TALP - Grup investigació UPC. GRAPH-MED 2019 [en línia]. [Consulta 25 març 2020]. Disponible a http://www.talp.upc.edu/project-detail/497/GRAPH-MED%20.

[3] Antonio Moreno - Instituto de Ingeniería del Conocimiento (IIC). Universidad Autónoma de Madrid [en línia]. [Consulta 25 març 2020]. Disponible a http://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/.

[4] American Medical Informatics Association [en línia]. [Consulta 2 abril 2020]. Disponible a https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168320/.

[5] NAACL (Noth American Chapter of the Association for Computational Linguistics) [en línia]. [Consulta 2 abril 2020]. Disponible a https://www.cs.york.ac.uk/semeval-2013/.

[6] Knowledge Leaning Project - University of Havana & University of Alicante[en línia]. [Consulta 2 abril 2019]. Disponible a https://knowledge-learning.github.io/ehealthkd-2019/.

[7] Cristopher D.Manning, Prabhakar Raghavan, Hinrich Schütze - Introduction to Information Retrieval. Cambridge University [en línia]. [Consulta 11 d'abril de 2020]. Disponible a https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

[8] John Lafferty, Andrew McCallum, Fernando C.N.Pereira - Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. University of Pennsylvania [en línia]. [Consulta 12 d'abril de 2020]. Disponible a https://repository.upenn.edu/cis_papers/159/.

[9] Ross Kindermann, J. Laurie Snell - Markov Random Fields and Their Applications. American Mathematical Society [en línia]. [Consulta 12 d'abril de 2020]. Disponible a http://www.cmap.polytechnique.fr/~rama/ehess/mrfbook.pdf.

[10] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Janvin - A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137-1155. 2003.

[11] Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, Jordi Armengol-Estapé - Medical Word Embeddings for Spanish: Development and Evaluation. [en línia] [consulta 12 d'abril de 2020]. Disponible a https://www.aclweb.org/anthology/W19-1916.pdf.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean - Efficient Estimation of Word Representations in Vector Space. [en línia] [consulta 12 d'abril de 2020]. Disponible a https://arxiv.org/abs/1301.3781.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov - Enriching Word Vectors with Subword Information. [en línia] [consulta 12 d'abril de 2020]. Disponible a https://arxiv.org/abs/1607.04606.

[14] Zhiheng Huang, Wei Xu, Kai Yu - Bidirectional LSTM-CRF Models for Sequence Tagging. Baidu research [en línia] [consulta 13 d'abril de 2020]. Disponible a https://arxiv.org/pdf/1508.01991v1.pdf.

[15] Nathan Greenberg, Trapit Bansal, Patrick Verga, Andrew McCallum - Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets. College of Information and Computer Sciences, University of Massachusetts Amherst [en línia] [consulta 13 d'abril de 2020]. Disponible a https://www.aclweb.org/anthology/D18-1306.pdf.

[16] Bill Y.Lin, Frank F.Xu, Zhiyi Luo, Kenny Q.Zhu - Multi-channel BiLSTM-CRF for Emerging Named Entity Recognition in Social Media. Shangai Jiao Ton University [en línia] [consulta 13 d'abril de 2020]. Disponible a https://www.aclweb.org/anthology/W17-4421.pdf.

[17] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer - Neural Architectures for Named Entity Recognition. Cargenie mellon University, NLP Group of Pompeu Fabra University [en línia] [consulta 13 d'abril de 2020]. Disponible a https://arxiv.org/pdf/1603.01360.pdf.

[18] Richa Chaturvedi, Deepak Arora, Pawan Singh - Conditional Random Fields and Regularization for efficient label prediction. ARPN Journal of Engineering and Applied Sciences [en línia] [consulta 14 d'abril de 2020]. Disponible a http://www.arpnjournals.org/jeas/research_papers/rp_2018/jeas_1018_7334.pdf.